Data Visualisation Report

By Lachlan Joshua McDonald

Introduction

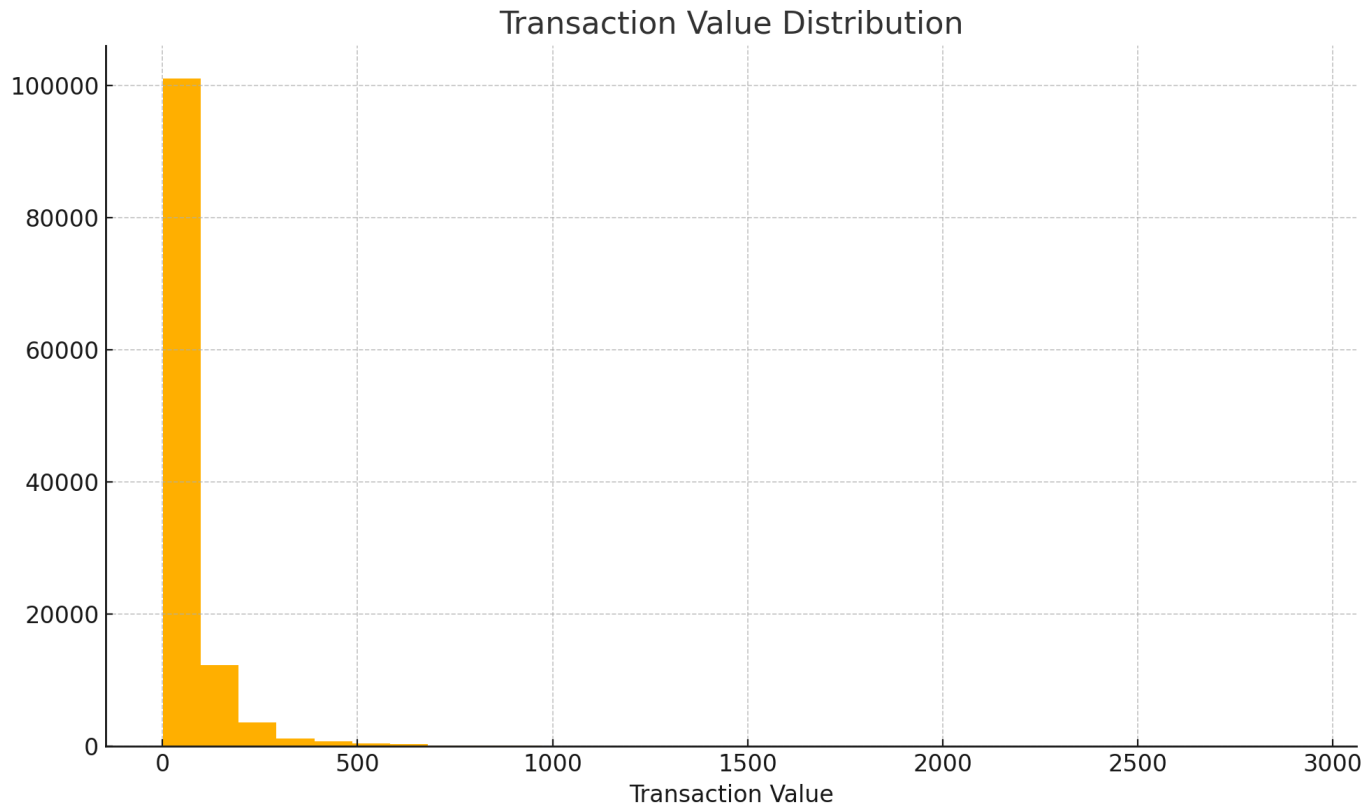
I selected the **"Synthetic Woolworths Cleaned"** spreadsheet from the synthetic data folder on GitHub due to its substantial size and the fact that it had already undergone a data cleaning process, making it an ideal candidate for analysis.

Each row in the dataset represents a unique transaction ID. To enhance the dataset, I added several new columns. One of these is **basket size**, which indicates how many units of a specific item were purchased in each line. By summing the quantity of all line-items per basket, I can better visualize purchase behavior and assess the number of items typically bought in a transaction.

Additionally, I converted the **purchase date** into a proper datetime format, which enables accurate temporal analysis. From this, I extracted features such as **week of the year**, **fortnight**, **calendar month**, and **quarter**. These time-based fields provide a strong foundation for identifying **seasonal trends** within the data.

Transaction Value

I chose to first start with a simple analysis. I focused on transaction value so that I could determine statistics about the value of transactions, such as where most transactions were happening.



Graph Analysis

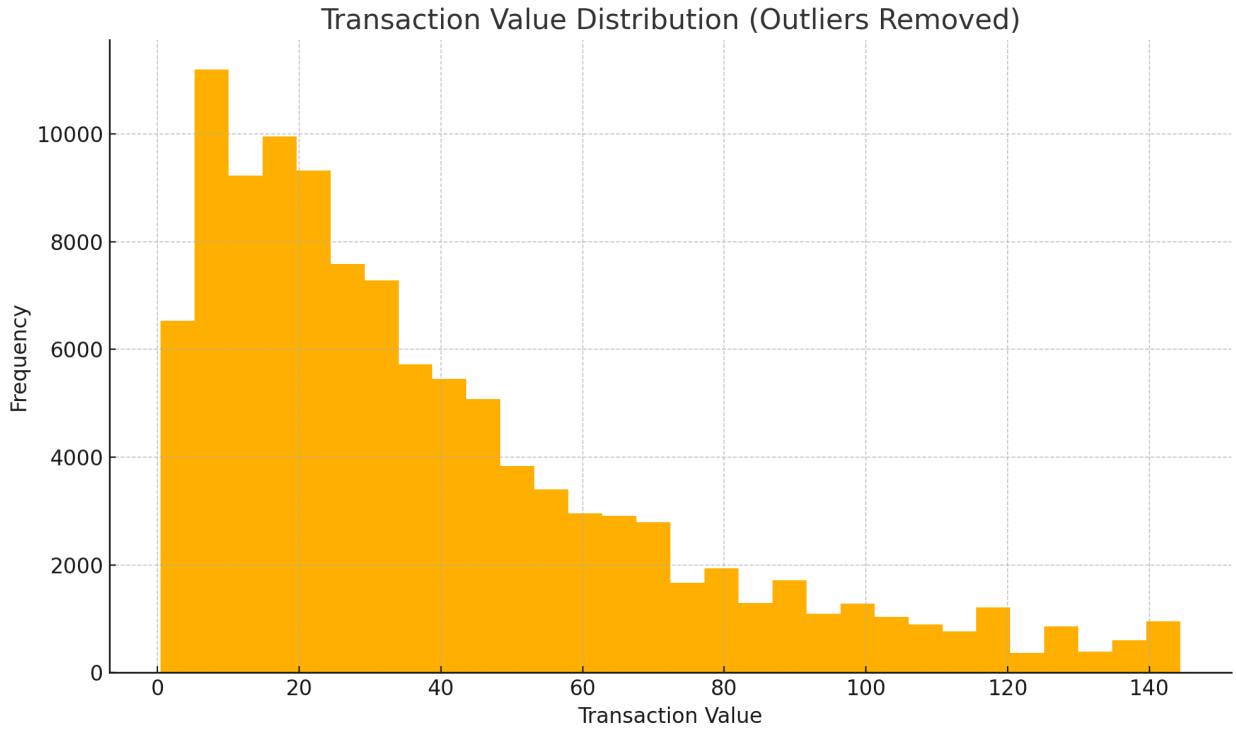
- The distribution is **heavily right-skewed**, with the vast majority of transactions concentrated below **$100**.
- A **long tail** extends toward **$3000**, caused by a small number of **extreme outliers**.
- The vertical spike near the origin suggests **most transactions are of low value**, consistent with typical grocery spending behavior.
- This skew heavily distorts the graph, making it difficult to interpret underlying trends without outlier removal.

This chart shows the **distribution of transaction values** across the dataset and was one of the most revealing visualizations in the early stages of analysis. It clearly demonstrates that most

transactions fall under **$100**, while a small number extend far beyond that, reaching up to **$2916**. These outliers distort the overall distribution, making it difficult to visually analyze the typical transaction range. To correct this, an outlier threshold was calculated using the **Q3 + 1.5 × IQR** method, which resulted in a cutoff of **$144.75**. This identified **10,926 transactions** as outliers. Removing these from the original **120,222 rows** left **109,296 more representative transactions**. Once outliers were excluded, the transaction value distribution became much easier to interpret, confirming the initial insight: **most transactions are small and routine**, characteristic of everyday grocery shopping patterns.

Removing Outliers

I decided to remove outliers from the transaction value distribution by using the **Q3 + 1.5 × IQR** method as stated above



Graph Analysis

- **Heavy left skew:** Over 50% of transactions fall below $30, and nearly 75% below $50, with frequency tapering off toward the $144.75 cap.
- **Primary mode around $10–$20:** The tallest bars sit at $10–$20, showing that most customers make small, routine grocery runs.
- **Long tail:** Even after outlier removal, there's a noticeable tail stretching from $50 to $144, indicating a smaller but steady stream of larger baskets.
- **Implications for reporting:** A log-scale or grouping into low-, mid-, and high-value tiers might make this distribution easier to compare month-to-month or segment by customer type.

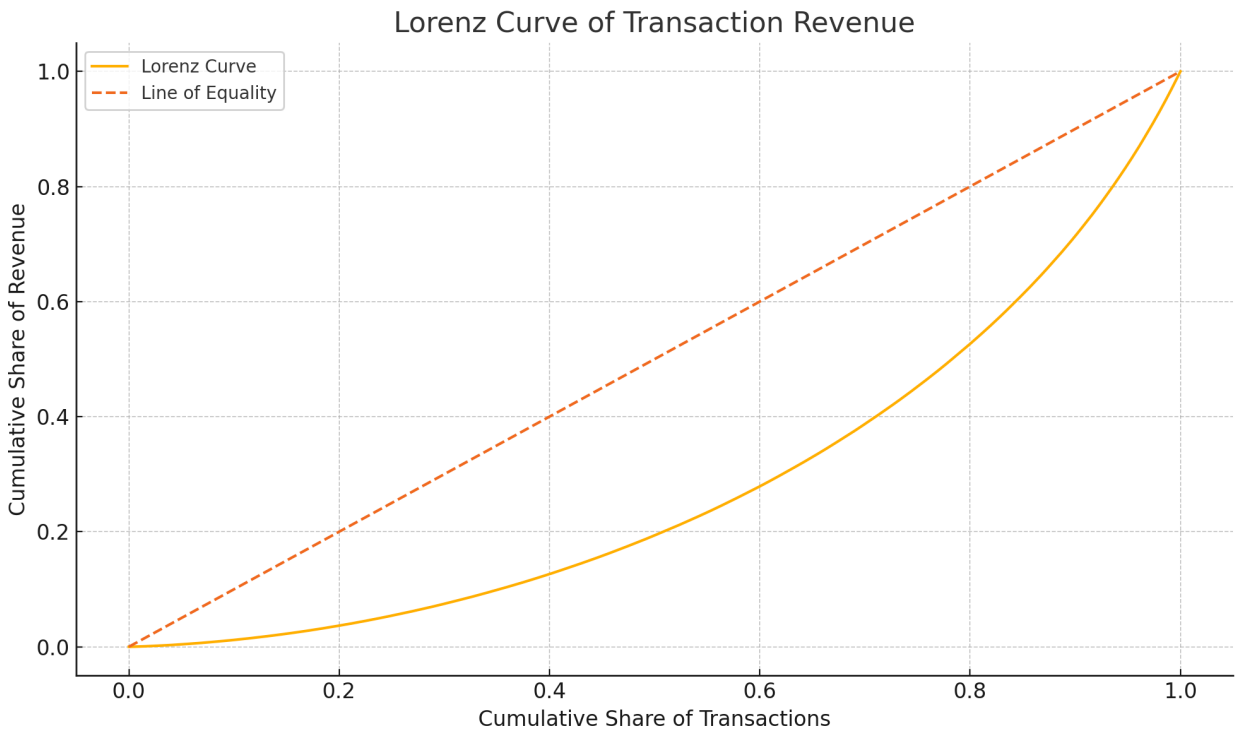
After excluding transactions above **$144.75**, the daily value distribution reveals a pronounced **right skew**: the **mean** is **$39.06** and the **median** is **$30**, confirming that most purchase amounts cluster tightly at the lower end. The **highest frequencies** appear in the **$10–$20** range, reflecting that routine, small-basket shopping dominates at Woolworths. A thinner but persistent tail remains between **$50** and **$144**, showing occasional mid- to high-value transactions.

Comparing the pre- and post-filter histograms clearly illustrates how just a few extreme values can stretch and mask the structure of the bulk of the data. For more nuanced insights, we could consider a log transformation or binning transactions into value brackets (e.g., $0–$30, $30–$60, $60–$144) to better highlight changes over time or differences across customer segments.

Statistic	**Before Removal**	**After Removal**
Mean	$62.28	$39.06
Median	$33	$30
Max	$2916	$144.40

Lorenz Curve of Transaction Revenue

I chose to further the analysis by using Lorenz Curve of Transaction Revenue. This technique is useful for determining the equality of transaction revenue. If it follows or closely follows the line of equality, then the data is the same or close to equal



Graph Analysis

- The Lorenz Curve **bows significantly below** the line of equality, indicating **high inequality** in transaction revenue.
- This means a **small percentage of transactions** are responsible for a **large share of total revenue**.
- Visually, it suggests that **perhaps the top 20% of transactions account for more than 60–70% of revenue**, a typical pattern in retail known as the **Pareto Principle** (80/20 rule).
- In practice, this implies that **Woolworths relies heavily on a small set of high-value transactions** to generate most of its revenue.

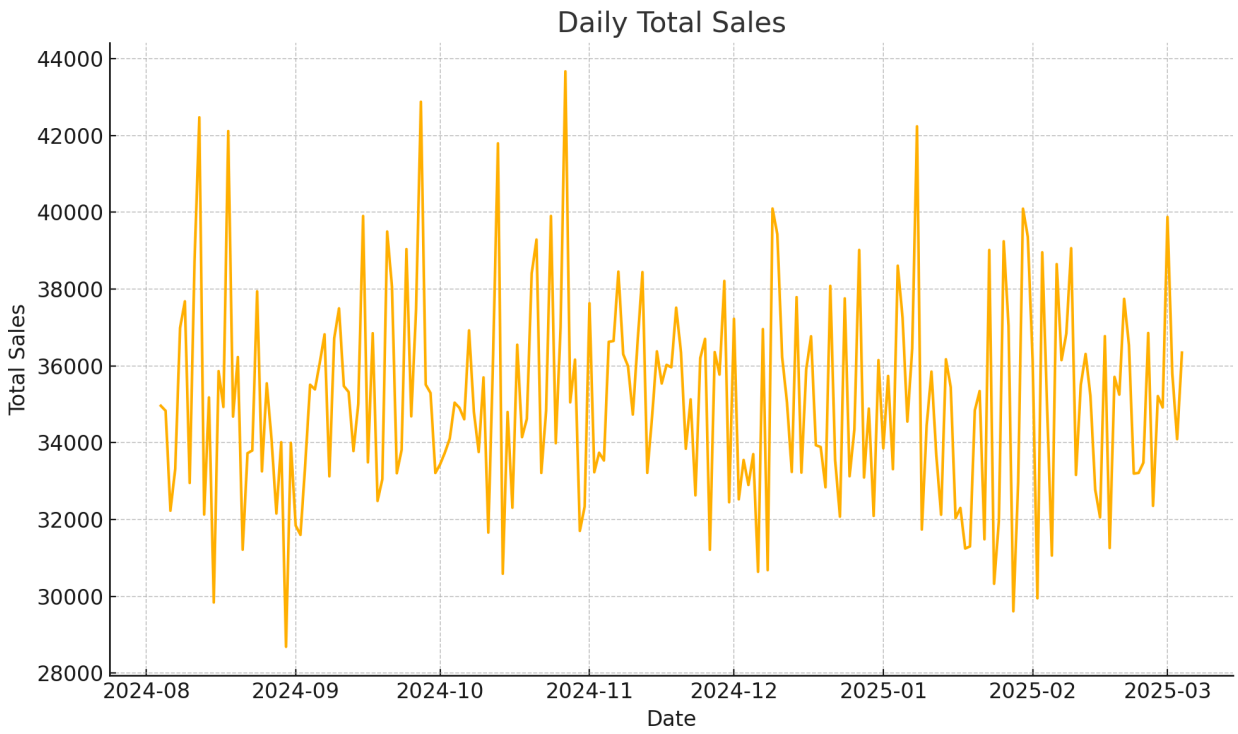
The Lorenz Curve of transaction revenue reveals a significant imbalance in how revenue is distributed across transactions. The sharp curve below the line of equality indicates that a small

proportion of transactions contributes disproportionately to total revenue. This suggests a high level of inequality, where the majority of revenue is driven by a minority of high-value transactions. Such a pattern is consistent with the Pareto Principle, where approximately 20% of transactions may account for around 80% of the revenue. This insight highlights the importance of understanding and targeting high-spending customers, as they play a crucial role in overall business performance.

Quantile	TransactionValue
0.25	$15.00
0.5	$30.00
0.75	$54.00
0.9	$88.00
0.95	$109.00
0.99	$144.00

Daily Total Sales

In order to determine if there is any seasonality, the actual daily total sales have to be observed first. The time period spans several months which gives enough range to



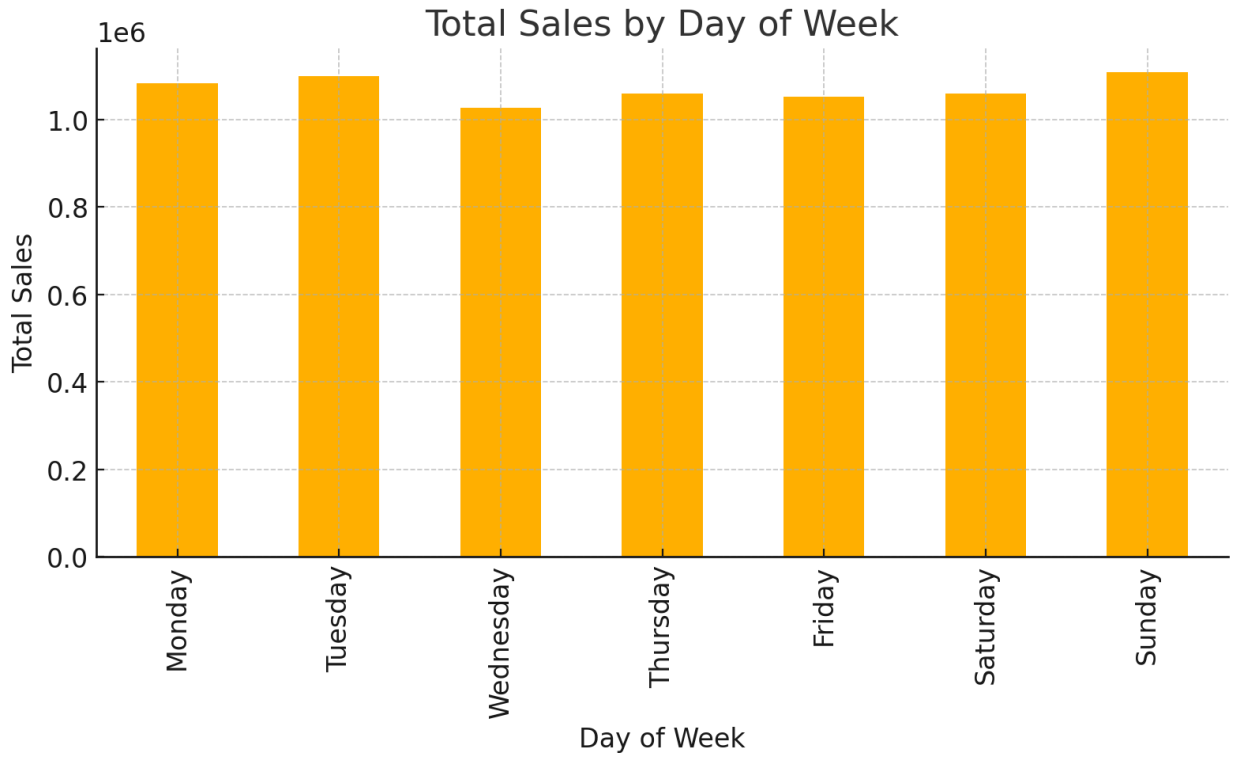
Graph Analysis

- Daily sales fluctuate between **$30,000 and $42,000**, with occasional spikes and dips.
- No clear **seasonal pattern** or **upward/downward trend** is immediately visible due to noise.
- The line shows **high volatility**, which can mask meaningful changes in business performance.
- Several abrupt **sales drops** may indicate data issues, holidays, or operational anomalies.
- A **moving average** or **monthly aggregation** could improve clarity and trend detection.

This line graph illustrates the **daily total sales** over a period spanning from August 2024 to March 2025. While it provides a general sense of the store's performance—showing that sales typically range between **$30,000 and $42,000 per day**, though the high degree of **daily fluctuation** makes it difficult to extract deeper insights in its current form. The graph does not clearly reveal **weekly patterns, seasonal trends, or anomalies**, due to the visual noise caused by the day-to-day volatility. Although it is useful for confirming that overall sales are relatively stable throughout the period, more meaningful patterns would emerge if the data were smoothed—such as by calculating a **7-day rolling average** or **aggregating by week or month**. This would allow for clearer trend analysis and make it easier to detect growth, dips, or seasonal effects.

Total Sales by Day of Week

I wanted to figure out more about total sales, and decided to look at different days of the week to see if any specific day had more of a chance of having more sales due to some assumptions relating to people being quite busy throughout the weekdays



Graph Analysis:

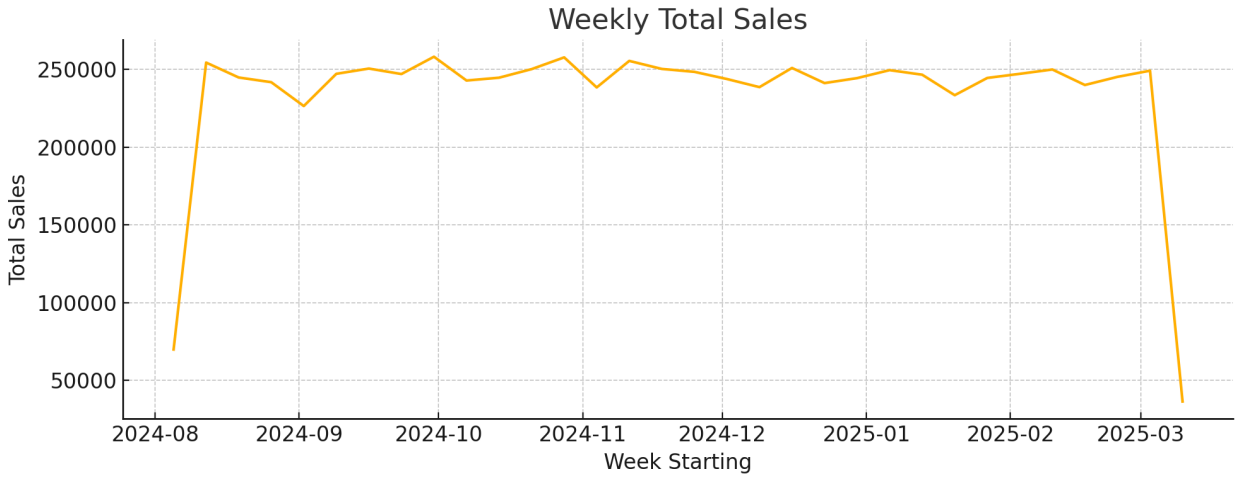
- **Sunday** and **Tuesday** show the **highest total sales**, peaking just above $1.1 million.
- **Wednesday** and **Friday** have slightly lower totals, though the drop is minor.
- Sales across all days are fairly **uniform**, with **low day-to-day variance**.
- This suggests **stable daily demand**, without strong reliance on specific peak days.
- May indicate that **customer visits are evenly distributed**, possibly due to everyday essentials and routine purchases.

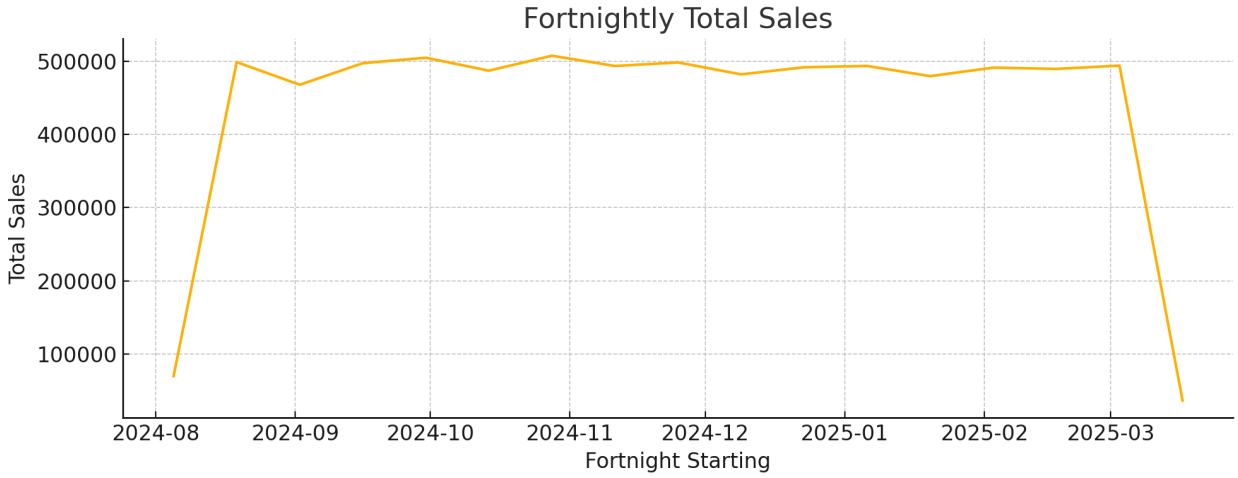
This bar chart presents the **total sales aggregated by day of the week**. From the visual, we can observe that **Sunday and Tuesday** have the highest total sales, while **Friday and Wednesday** show the lowest, though the differences are relatively small. Overall, the values across all days are quite close, suggesting that **Woolworths experiences steady consumer activity throughout the week**. The consistency implies a balanced flow of customers, likely

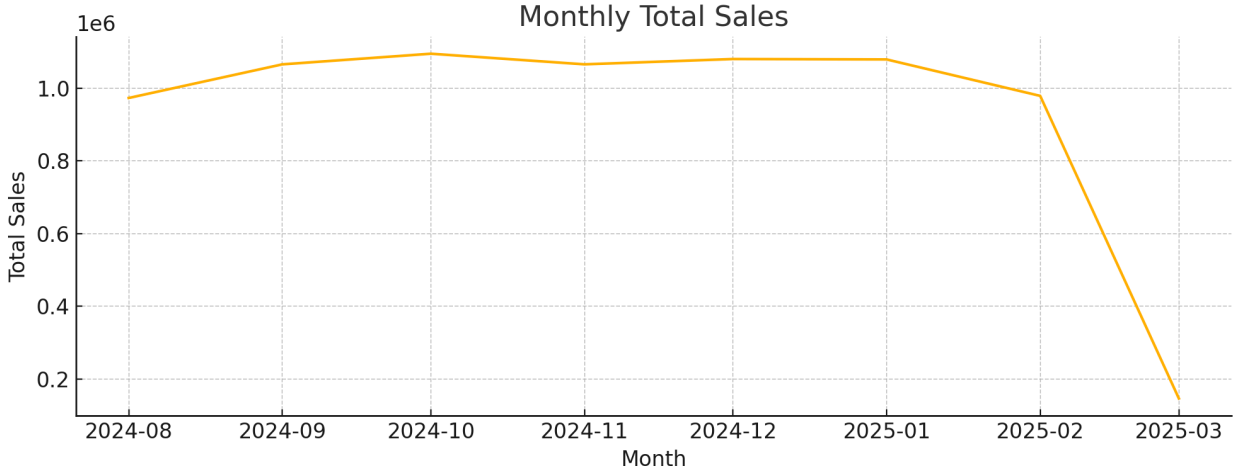
due to regular grocery habits rather than specific shopping days. However, the slight edge on weekends and early-week days could be indicative of restocking behaviors after the weekend or ahead of the workweek.

Seasonal Total Sales

To begin directly looking at seasonal data, I began by taking 3 kinds of measurements: Weekly, Fortnightly, and Monthly. The reason for these 3 choices is that this data set was very small, only spanning several months, so anything larger than a month would not be particularly useful







Graph Analysis:

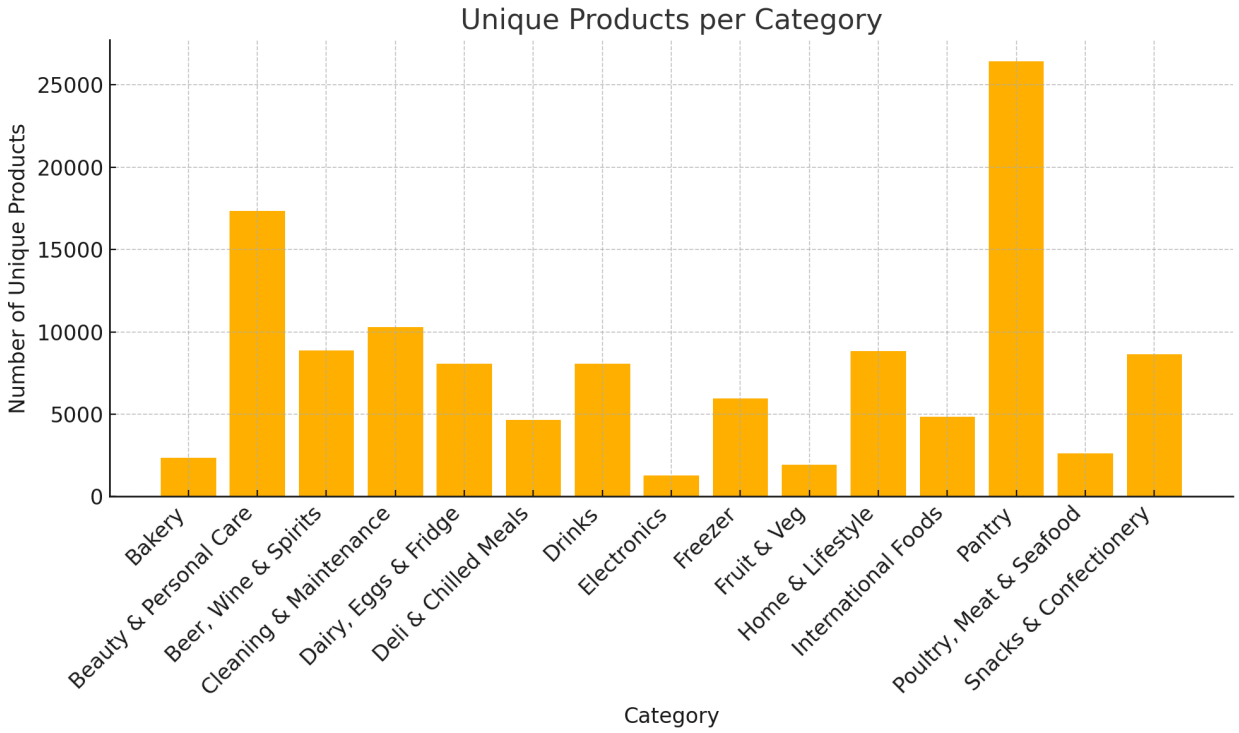
- **Weekly Sales**:
    - Fluctuate slightly between $230,000–$260,000.
    - Most weeks maintain a steady volume.
    - Sharp dip at the start and end due to partial weeks.
- **Fortnightly Sales**:
    - Hover around $480,000–$510,000 per two-week period.
    - Demonstrates better smoothing and less noise than weekly data.
    - Final point is clearly incomplete.
- **Monthly Sales**:
    - Range from just under $1 million (August) to about $1.1 million (October–January).
    - October and December show **minor peaks**, likely tied to seasonal/holiday demand.
    - March drop is a clear result of partial data, not actual performance decline.
- Across all three:
    - **No major trend of growth or decline**—sales are **consistently strong and stable**.
    - Graphs are useful for **high-level reporting**, budgeting, and capacity planning.
    - Aggregation helps identify **seasonal boosts** and filter out daily anomalies.

These three line graphs provide a **multi-resolution view of total sales over time**, from weekly to fortnightly to monthly aggregation. The overall trend is remarkably consistent across all three timeframes, with sales typically remaining **stable throughout the observed period (Aug 2024 – Feb 2025)**. However, all three graphs show **sharp drops** at the beginning and end of the period due to **incomplete data collection** in those boundary months rather than actual business performance dips. Aggregating the data this way removes much of the noise seen in

the daily sales chart and helps reveal a **steady baseline of customer spending**. Notably, monthly totals show **slight seasonal increases in October and December**, which may reflect pre-holiday shopping behavior.

Products per Category

Seasonality was hard to determine by checking all categories of products, so I decided it would be best to focus on one product category. The reason only one is necessary is due to the fact that this is synthetic data, so seeing many graphs with data being almost entirely the same.



Graph Analysis

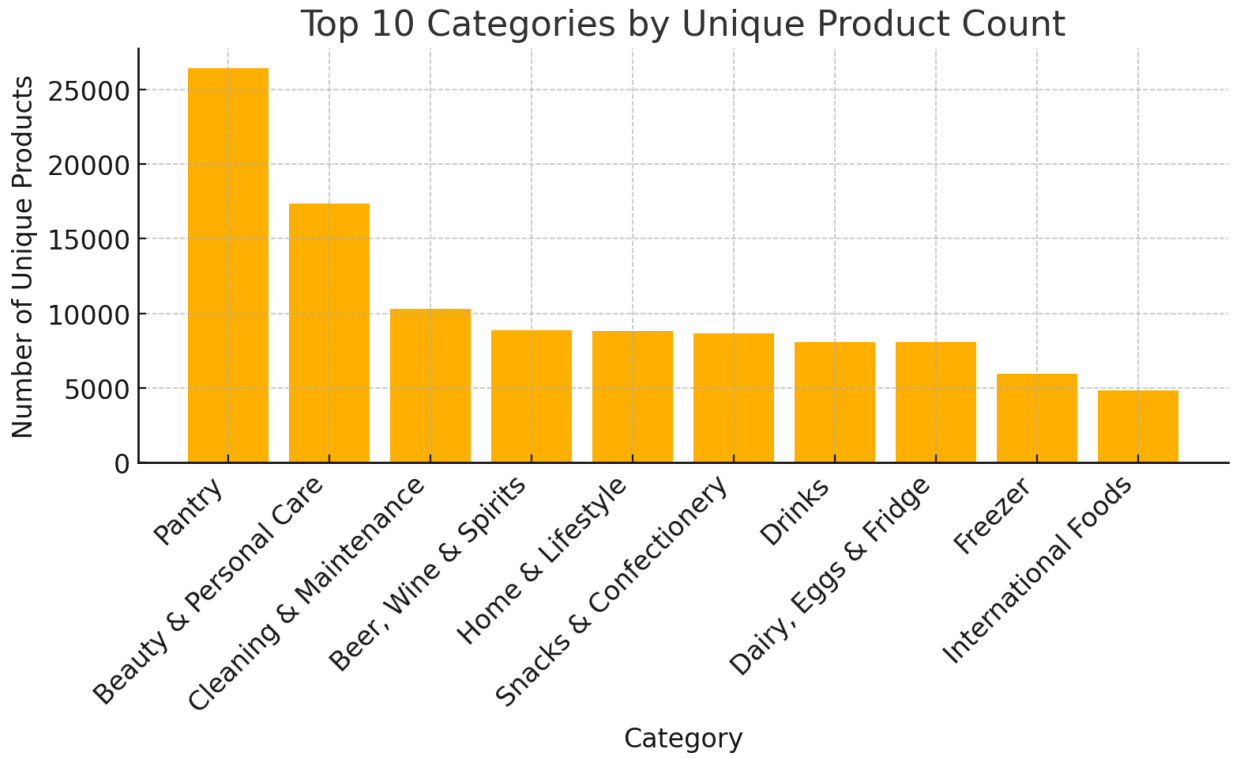
- **Pantry** stands out with over **25,000 unique products**, more than double any other category.
- **Beauty & Personal Care** and **Beer, Wine & Spirits** follow, each with high product variety around **17,000** and **10,000+**, respectively.
- Categories such as **Bakery**, **Poultry, Meat & Seafood**, and **Electronics** have the fewest unique items, all under **5,000**.
- **International Foods**, **Cleaning & Maintenance**, and **Snacks & Confectionery** show **moderate diversity**, with 8,000–9,000 unique products.
- The data suggests a strong emphasis on **shelf-stable, everyday goods** (like pantry items) rather than perishable or specialty items.

This bar chart illustrates the number of unique products offered per category, providing insight into the diversity of each department's range. The Pantry category dominates the chart with over

25,000 unique items, far outpacing every other category. This aligns with Pantry's role as a central grocery category that includes a vast array of shelf-stable goods. Following that, **Beauty & Personal Care** and **Beer, Wine & Spirits** also show high diversity, likely reflecting consumer preference for variety and brand differentiation in these spaces. In contrast, categories like **Bakery**, **Poultry, Meat & Seafood**, and **Electronics** show much lower counts, which makes sense given their perishability, storage constraints, or lower purchase frequency. The product distribution emphasizes that the dataset is structured around high-volume, high-rotation categories typical of general supermarkets.

Top 10 Categories by Unique Product Count

I organised the data to give the top 10 categories, to see which category would be the most important to begin looking for seasonality in



Graph Analysis:

- **Pantry** has the highest number of unique products (>25,000), dominating all other categories.
- **Beer, Wine & Spirits** and **Beauty & Personal Care** follow, each with ~17,000–18,000 items.
- **Electronics**, **Bakery**, and **Poultry, Meat & Seafood** have the **lowest product variety** (<4,000).
- Moderate variety observed in **Cleaning & Maintenance**, **Dairy, Eggs & Fridge**, and **International Foods** (~8,000–10,000).

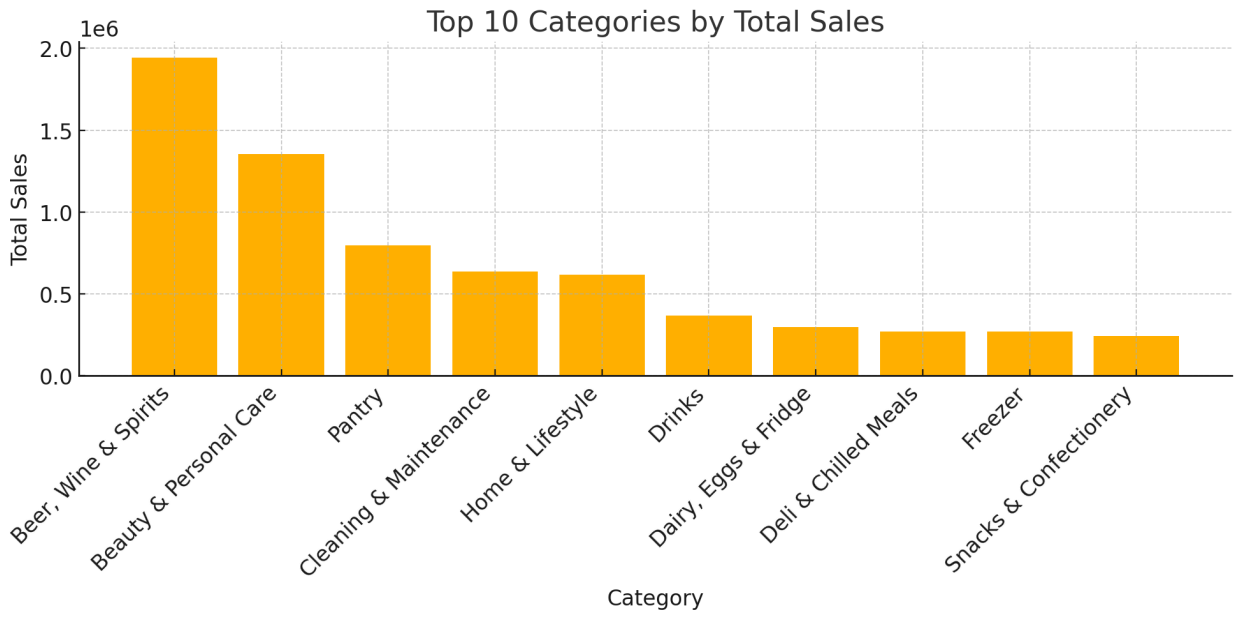
- Categories with high variety likely correspond to **broad subcategories or consumer demand for options**.

- Lower-variety categories may indicate **limited demand, perishability**, or **space constraints**.

This bar chart displays the **number of unique products available in each product category**. The most notable observation is that the **Pantry** category stands out significantly, with over **25,000 unique products**, far exceeding all other categories. This is followed by **Beer, Wine & Spirits** and **Beauty & Personal Care**, both offering high product diversity in the range of **17,000–18,000 items**. On the other end of the spectrum, **Electronics**, **Bakery**, and **Poultry, Meat & Seafood** have relatively low product variety. These differences likely reflect the nature of product turnover and shelf space allocation—Pantry items are typically dry goods with long shelf lives, allowing for greater variety, while fresh or highly specialized categories tend to have fewer distinct items. Understanding product variety is essential for inventory planning, marketing strategy, and optimizing customer choice.

Rank	Category	Unique Products
1	Pantry	26430
2	Beauty & Personal Care	17334
3	Cleaning & Maintenance	10306
4	Beer, Wine & Spirits	8860
5	Home & Lifestyle	8825
6	Snacks & Confectionery	8645
7	Drinks	8082
8	Dairy, Eggs, & Fridge	8066
9	Freezer	5951
10	International Foods	4849

Top 10 Categories by Total Sales



Graph Analysis:

- **Beer, Wine & Spirits** is the top-performing category in terms of sales (~$2M).
- **Beauty & Personal Care** and **Pantry** follow, both exceeding **$1M**.
- **Pantry**, despite having the most unique products, ranks **third in sales**—suggesting lower sales per item.
- **Cleaning & Maintenance** and **Home & Lifestyle** form a solid mid-tier in sales.
- **Freezer** and **Snacks & Confectionery** rank lowest in the top 10, each under **$300K**.
- A focused product mix (like in Beer & Wine) may drive **higher efficiency and profitability per item**.

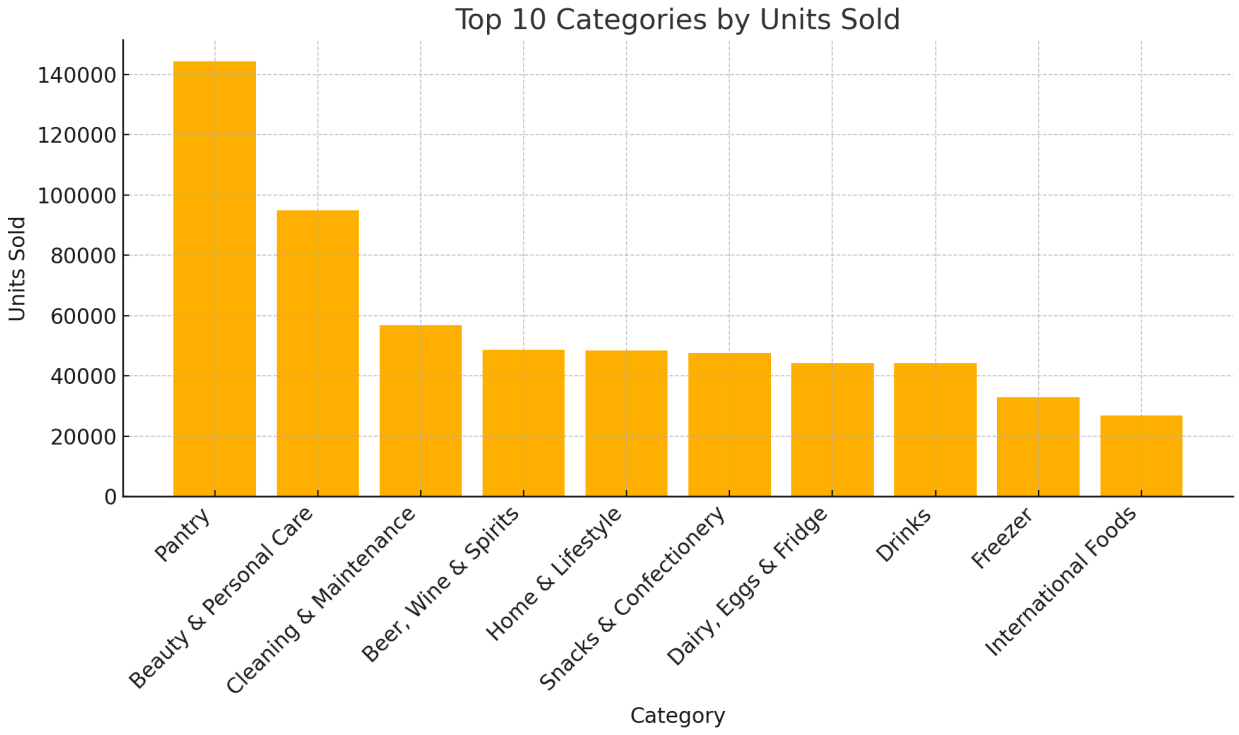
This bar chart highlights the **top 10 product categories by total sales**, revealing which areas contribute most significantly to overall revenue. **Beer, Wine & Spirits** leads by a large margin, generating close to **$2 million** in sales, followed by **Beauty & Personal Care** and **Pantry**, both with over **$1 million**. The remaining categories, including **Cleaning & Maintenance**, **Home & Lifestyle**, and **Drinks**, contribute moderately, while **Snacks & Confectionery** and **Freezer** categories appear at the lower end of the top 10. Interestingly, while **Pantry** had the highest number of unique products in the previous chart, it ranks only third in total sales. This indicates

that **high product variety does not always equate to the highest revenue**, and more focused categories like Beer & Wine may drive stronger per-product performance.

Category	TransactionCount	UnitsSold	TotalSales
Beer, Wine & Spirits	8,860	48,657	$1,943,973
Beauty & Personal Care	17,334	94,851	$1,353,582
Pantry	26,430	144,323	$797,173
Cleaning & Maintenance	10,306	56,767	$635,436
Home & Lifestyle	8,825	48,377	$617,247
Drinks	8,082	44,175	$366,807
Dairy, Eggs & Fridge	8,066	44,253	$298,056
Deli & Chilled Meals	4,662	25,599	$271,773
Freezer	5,951	33,019	$269,621
Snacks & Confectionery	8,645	47,703	$244,857
Poultry, Meat & Seafood	2,634	14,612	$219,071.88
Electronics	1,267	7,012	$217,978.50
International Foods	4,849	26,923	$125,522.44
Bakery	2,373	13,190	$77,018.14
Fruit & Veg	1,938	10,733	$48,710.74

Categories by Units Sold

Category	TransactionCount	UnitsSold	TotalSales
Pantry	26,430	144,323	$797,173
Beauty & Personal Care	17,334	94,851	$1,353,582
Cleaning & Maintenance	10,306	56,767	$635,436
Beer, Wine & Spirits	8,860	48,657	$1,943,973
Home & Lifestyle	8,825	48,377	$617,247
Snacks & Confectionery	8,645	47,703	$244,857
Dairy, Eggs & Fridge	8,066	44,253	$298,056
Drinks	8,082	44,175	$366,807
Freezer	5,951	33,019	$269,621
International Foods	4,849	26,923	125,522.44
Deli & Chilled Meals	4,662	25,599	$271,773
Poultry, Meat & Seafood	2,634	14,612	219,071.88
Bakery	2,373	13,190	77,018.14
Fruit & Veg	1,938	10,733	48,710.74
Electronics	1,267	7,012	217,978.50



Graph Analysis:

- **Pantry** leads significantly with **~140,000+ units sold**, confirming its central role in household shopping.

- **Beauty & Personal Care** and **Cleaning & Maintenance** follow with strong sales in **non-food essentials**.

- **Beer, Wine & Spirits** sells fewer units but generates **higher revenue**, indicating **premium pricing**.

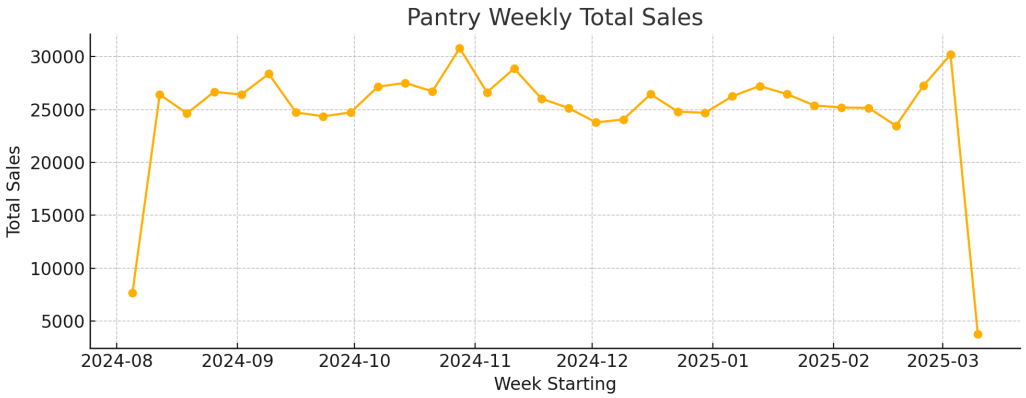
- **Snacks & Confectionery**, **Dairy**, and **Drinks** show consistent unit sales across smaller, everyday items.

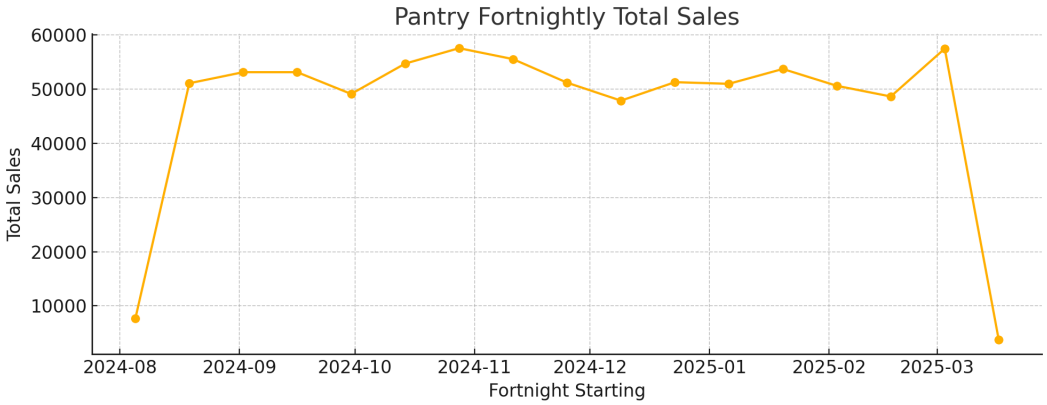
- **Freezer** and **International Foods** appear at the lower end, likely due to **limited variety or niche demand**.

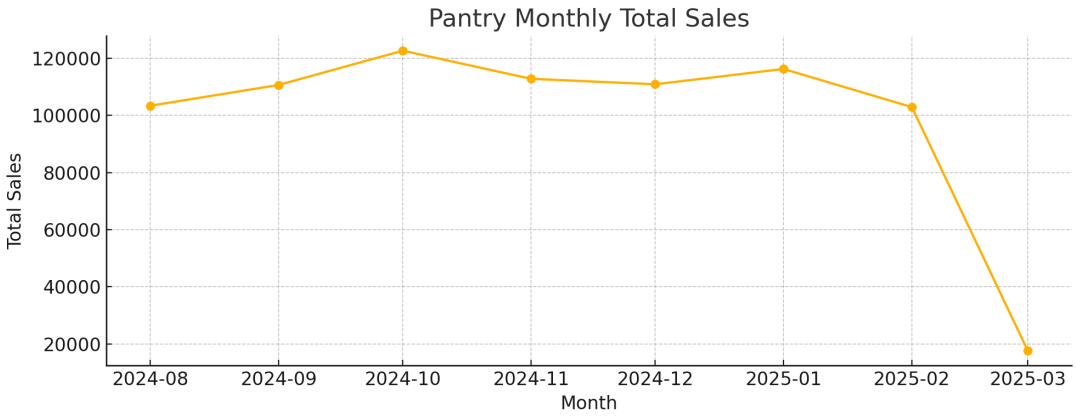
- Overall, this graph emphasizes **volume-driven categories**, while previous sales revenue charts highlight **value-driven categories**.

This bar chart highlights the top 10 product categories based on **total units sold**, providing insight into customer purchasing behavior by volume. **Pantry** dominates the chart, with over **140,000 units sold**, far surpassing all other categories. This aligns with its high variety and everyday necessity, indicating frequent and consistent purchases. **Beauty & Personal Care** and **Cleaning & Maintenance** follow, each selling between **90,000–100,000 units**, showing strong demand for non-food essentials. Interestingly, while **Beer, Wine & Spirits** ranked first in total sales revenue, it ranked only fourth in units sold, suggesting a **higher average price per unit**. In contrast, categories like **Snacks & Confectionery** and **Dairy, Eggs & Fridge** maintain solid sales volumes with likely lower price points. Overall, the chart reveals that **high unit sales do not always equate to high revenue**, and understanding both metrics is key to effective product and pricing strategy.

Pantry Seasonal Sales







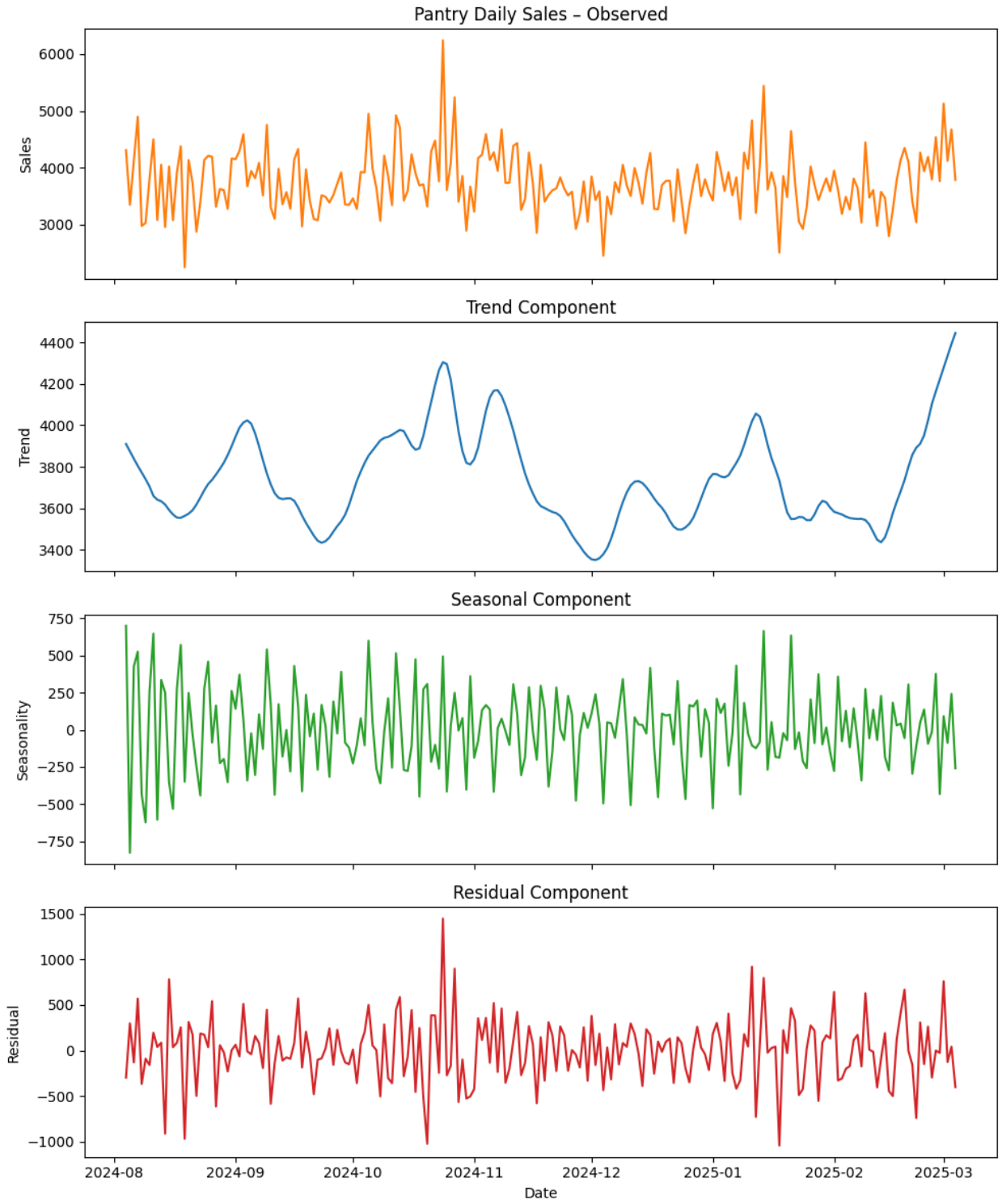
Graph Analysis:

- **Weekly Sales**:
    - Mostly consistent between **$24,000 and $28,000** per week.

    - **Spikes** around **mid-November** and **early March**, possibly due to **Black Friday promotions** and **early autumn sales**.
    - Initial and final weeks show **abnormally low values**, likely due to **incomplete data**.
- **Fortnightly Sales**:
    - Reinforces the same pattern with clearer **peaks in Nov and March**.
    - Smoother trend with less noise compared to weekly data.
    - Last fortnight also shows a **sharp drop**, confirming incomplete data collection.
- **Monthly Sales**:
    - Displays a steady trend from **August 2024 to February 2025**, peaking in **October** and **January** (~$120,000).
    - The **March value drops drastically**, which is misleading due to **missing sales data** for the rest of the month.
    - Monthly aggregation smooths out smaller spikes visible in weekly/fortnightly views.

The three graphs showing **Pantry sales at weekly, fortnightly, and monthly levels** collectively reveal a stable pattern with occasional fluctuations. Weekly sales typically range between **$24,000 and $28,000**, indicating consistent demand. Notable peaks appear around **mid-November**—likely corresponding with **Black Friday promotions or holiday stocking**—and again in **early March**, possibly reflecting early autumn shopping behavior or promotional events. These trends are echoed more clearly in the **fortnightly view**, which reduces daily noise and highlights cyclical movement. The **monthly graph**, while effective for visualizing broader trends, smooths over these short-term spikes and is impacted by **incomplete data** in March, leading to an artificially low value. Overall, the data suggests **moderate seasonal variation** in Pantry sales, particularly in the **lead-up to holidays** and **new year resets**, which is valuable for inventory planning and promotional timing.

STL Decomposition

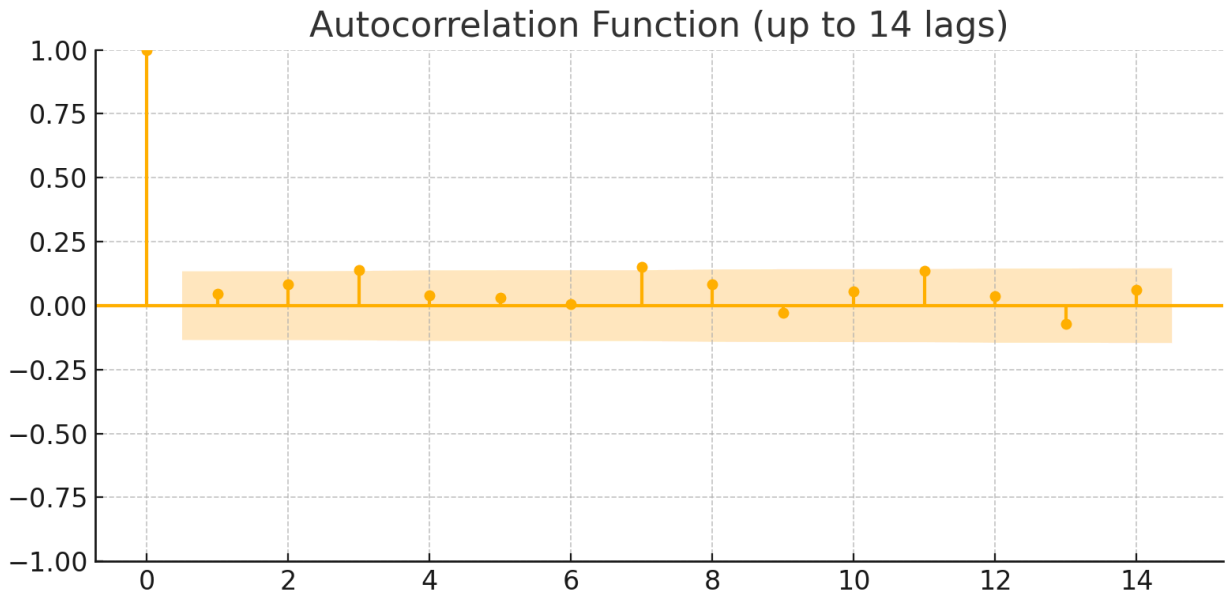


Graph Analysis:

- **Observed Sales**:
    - Fluctuate daily between ~$3,000 and $4,500.
    - Notable spikes in **late October** and **early March** (~$6,000+).
    - Overall very noisy and hard to interpret directly.
- **Trend Component**:
    - Smooth upward waves in **September**, **November**, and **late February**.
    - General long-term trend is **positive/stable**.
- **Seasonal Component**:
    - Small oscillations (±600), **no strong weekly/monthly cycle**.
    - Implies weak or inconsistent seasonality in pantry sales.
- **Residual Component**:
    - High variability; spikes align with observed anomalies.
    - Suggests **sales are affected by unpredictable events** (e.g. sales, holidays, supply/demand shocks).

These four graphs show the decomposition of daily pantry sales into observed, trend, seasonal, and residual components. The **observed data** is highly volatile, with daily sales typically ranging between **$3,000 and $4,500**, and occasional spikes above **$6,000**, notably around **late October** and **early March**, likely tied to promotions or holidays. The **trend component** smooths this noise and shows clear rises in **September**, **November**, and again in **late February**, suggesting potential sales growth heading into March. The **seasonal component** fluctuates within a ±600 range but shows **no strong recurring pattern**, indicating that pantry sales don't follow a rigid seasonal cycle. The **residual graph** reveals significant short-term randomness, especially around spikes in the observed chart, which reinforces the idea that pantry sales are driven more by **sporadic events or campaigns** than by predictable seasonality. Overall, sales are stable with a mild upward trend and influenced mostly by irregular, one-off factors.

Autocorrelation Function

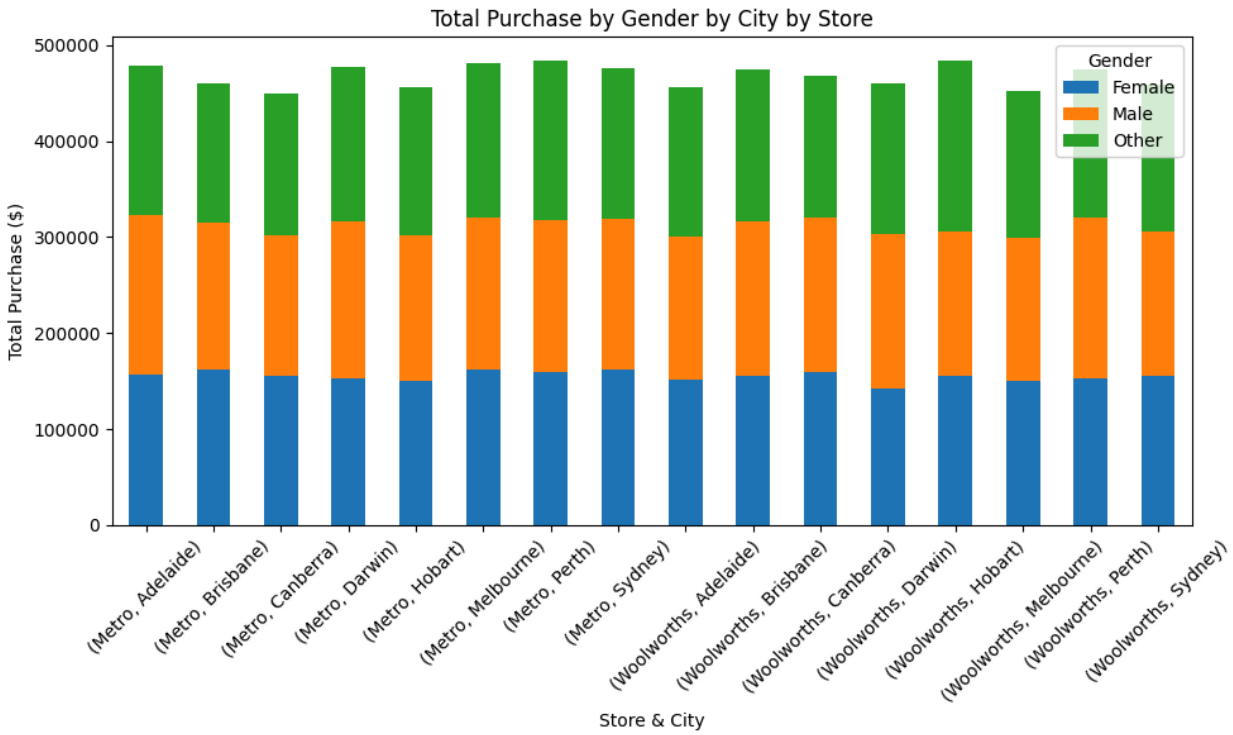


Graph Analysis:

- **Lag 7** shows a **positive autocorrelation of 0.161**, slightly outside the confidence band.
- Indicates a **weekly cycle**, where today's sales are modestly correlated with sales a week ago.
- Other lags show weak or insignificant autocorrelation—**no strong daily persistence**.
- **Fourier regression cosine term** (period = 7 days) supports this with a **positive coefficient (~94)** and **p = 0.075**.
- Suggests a **mild, noise-obscured weekly seasonality** in pantry sales.

This autocorrelation plot examines the relationship between pantry sales over time, with lags up to 14 days. Most autocorrelations are small and fall within the 95% confidence interval, indicating low persistence in day-to-day fluctuations. However, the **spike at lag 7**, with a value of **0.161**, stands out as it slightly exceeds the upper bound of the confidence band. This suggests a **modest weekly pattern**, where sales on a given day tend to be positively correlated with those exactly one week earlier. This finding is supported by Fourier regression, where the **cosine term with a 7-day period** has a **positive coefficient (~94)** and a **p-value of 0.075**. Although this p-value does not meet conventional significance thresholds (e.g. $p < 0.05$), it does indicate a **weak but consistent weekly seasonality** amid high daily noise. Overall, the synthetic data appears to reflect subtle weekly sales cycles, even if daily volatility makes them harder to detect with high confidence.

Total Purchase by Gender by City by Store



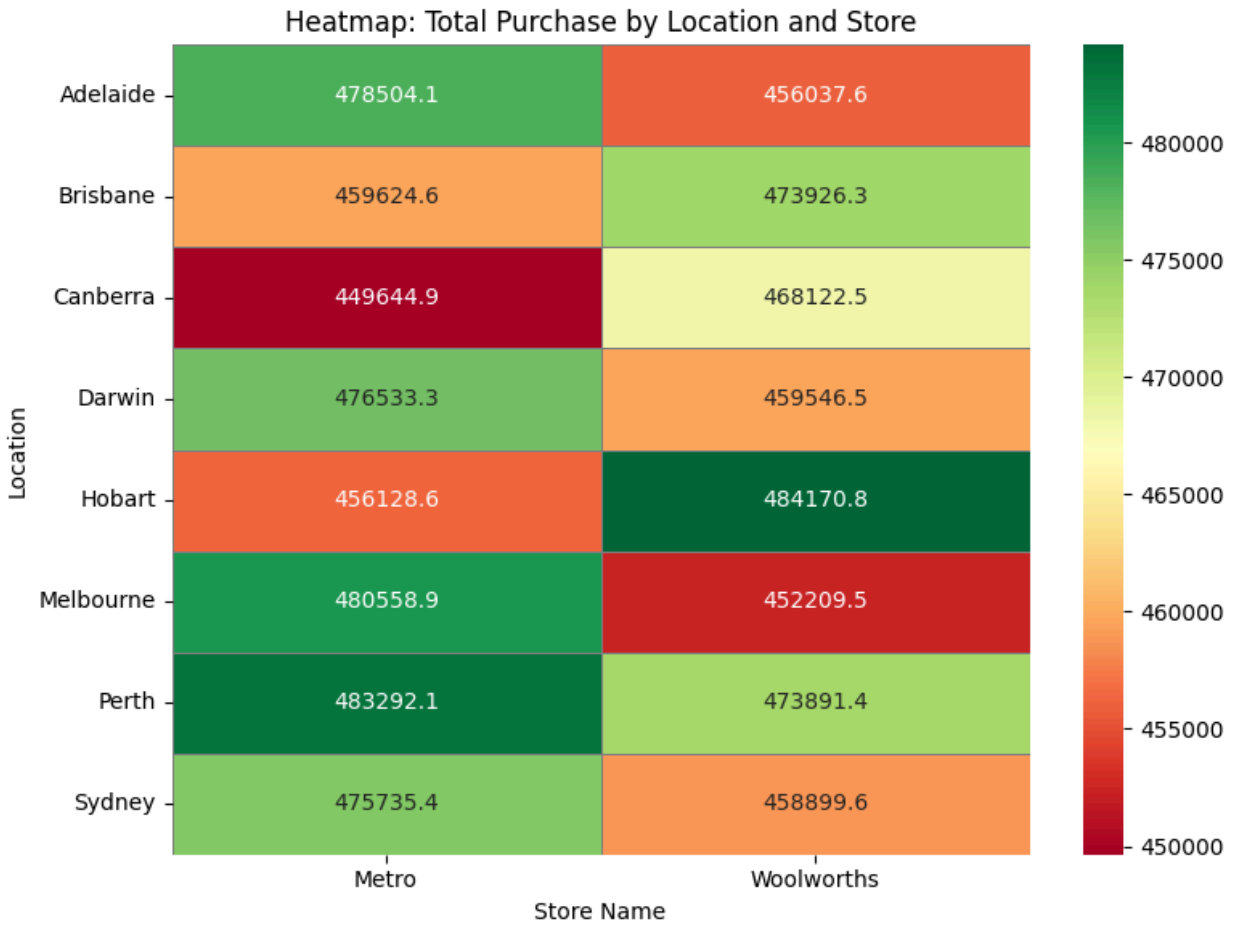
Graph Analysis:

- Every store-city combination shows **nearly identical totals across Female, Male, and Other**.
- **No visible demographic trend** by city or store type (e.g., Metro vs Woolworths).
- Implies that gender assignment in the dataset is **artificially balanced**, not naturally occurring.
- The **"Other" category** is unusually high, suggesting inclusion of non-disclosed or randomly generated entries.
- Demonstrates a key flaw of synthetic datasets: **unrealistic demographic distributions**, which limit real-world applicability.

This stacked bar chart shows the **total purchase amounts by gender across major Australian cities**, split by store type (**Metro vs Woolworths**). Visually, each location presents a **near-perfect three-way split** between **Female**, **Male**, and **Other** gender categories, with no meaningful variation between cities or store types. This uniformity is highly unusual and **not**

representative of real-world demographics, where gender distributions typically vary and non-binary or undisclosed gender categories rarely account for one-third of total customers. The likely explanation is that this pattern is **artificially imposed by the synthetic data generation process**, which seems to have evenly distributed purchases across gender for neutrality or balance. While this flaw limits real-world insight, it does serve as a good **reminder of synthetic data limitations**—particularly when analyzing demographic or behavioral variables. Nonetheless, it can still be used to **stress-test dashboards or analytical methods**, so long as its unrealistic structure is acknowledged.

Heat Map of Total Purchase by Location and Store



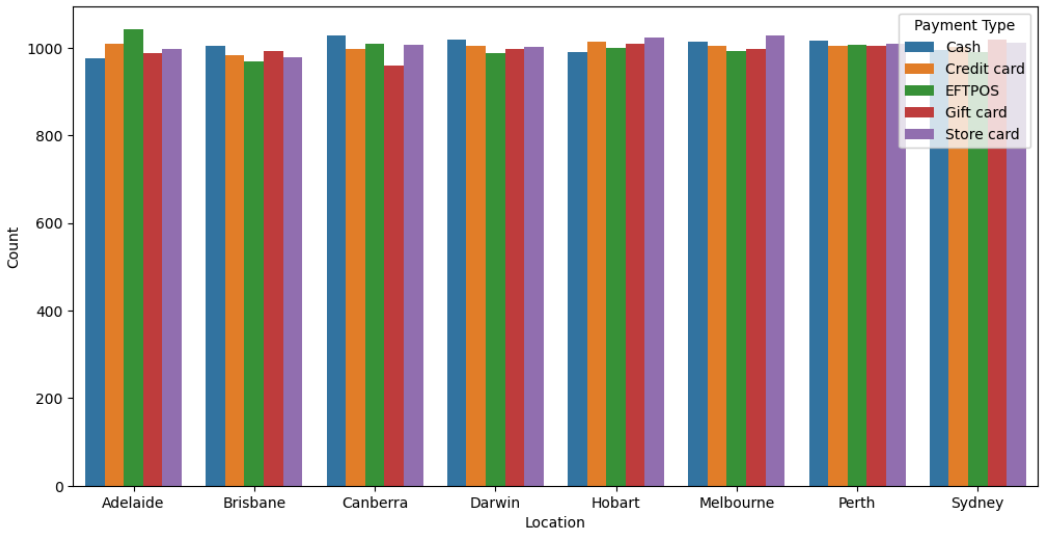
Graph Analysis:

- **Highest total purchase**: Hobart Woolworths – **$484,170.80**

- **Lowest total purchase**: Canberra Metro – **$449,644.90**
- Most cities show **less than 5% difference** between Metro and Woolworths stores.
- No single store or location dramatically outperforms others.
- Reinforces the idea of **even distribution in synthetic data**, likely for fairness or simplicity.
- Lacks real-world noise or skew caused by **urban density, income level, or local habits**.

This heat map displays the **total purchase amounts by location and store type** (Metro vs Woolworths). The values range from a **low of $449,644.90** (Canberra Metro) to a **high of $484,170.80** (Hobart Woolworths), indicating **very little variation** in total sales across the dataset. The narrow spread in values—less than a 10% difference between the highest and lowest—suggests that the synthetic data was **generated with balanced totals across all city-store combinations**. While real-world retail performance typically varies more significantly between locations due to demographics, foot traffic, and competition, this synthetic dataset shows a **uniformity** more characteristic of artificial generation rather than organic business trends. This uniformity may make it **less useful for exploring natural commercial differences**, but it remains valuable for testing visualizations and models under controlled conditions.

Count of Transaction by Payment Type per Location

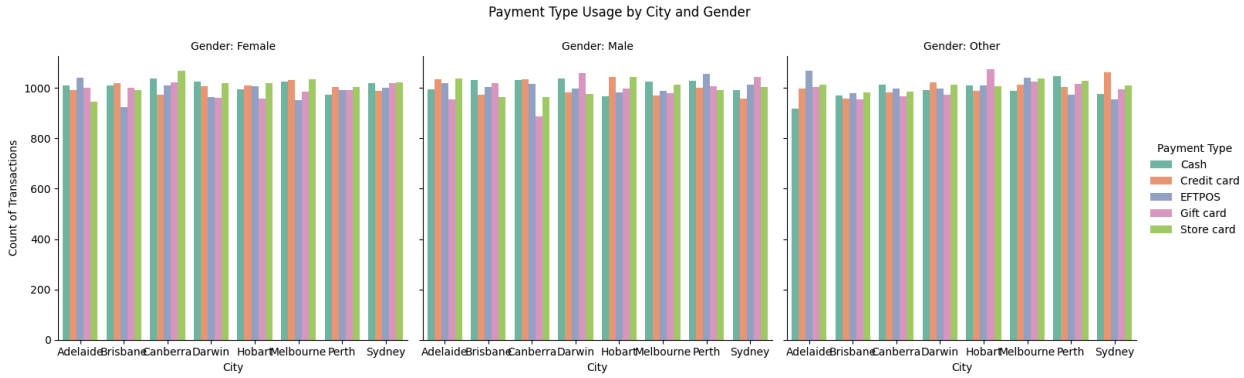


Graph Analysis:

- All five payment types appear in **nearly equal proportions** across every city.
- **No city-specific payment trends** are visible—e.g., no card-heavy or cash-dominant areas.
- Suggests synthetic logic applied **uniform random selection** to assign payment types.
- Contrasts with real-world data, where **EFTPOS or credit card dominance** is often observed.
- Highlights a key limitation of synthetic data for **payment behavior studies** or consumer profiling.

This bar chart displays the **count of transactions by payment type across different Australian cities**. Each city features nearly identical heights for all five payment methods: **Cash, Credit Card, EFTPOS, Gift Card, and Store Card** with minimal variation across locations. This uniformity is **highly unrealistic** in a real-world setting, where payment preferences typically differ by **region, age group, store type, and socioeconomic factors**. The near-equal distribution of payment types across all cities indicates that the **synthetic data has assigned payment methods with roughly equal probability**, regardless of location context. While this may simplify dataset generation, it renders the data **unsuitable for any serious behavioral or demographic payment analysis**. As a result, this chart serves as a strong example of how **synthetic datasets can obscure real-world patterns** and introduce biases that would not exist in naturally collected data.

Payment Type Usage By City and Gender

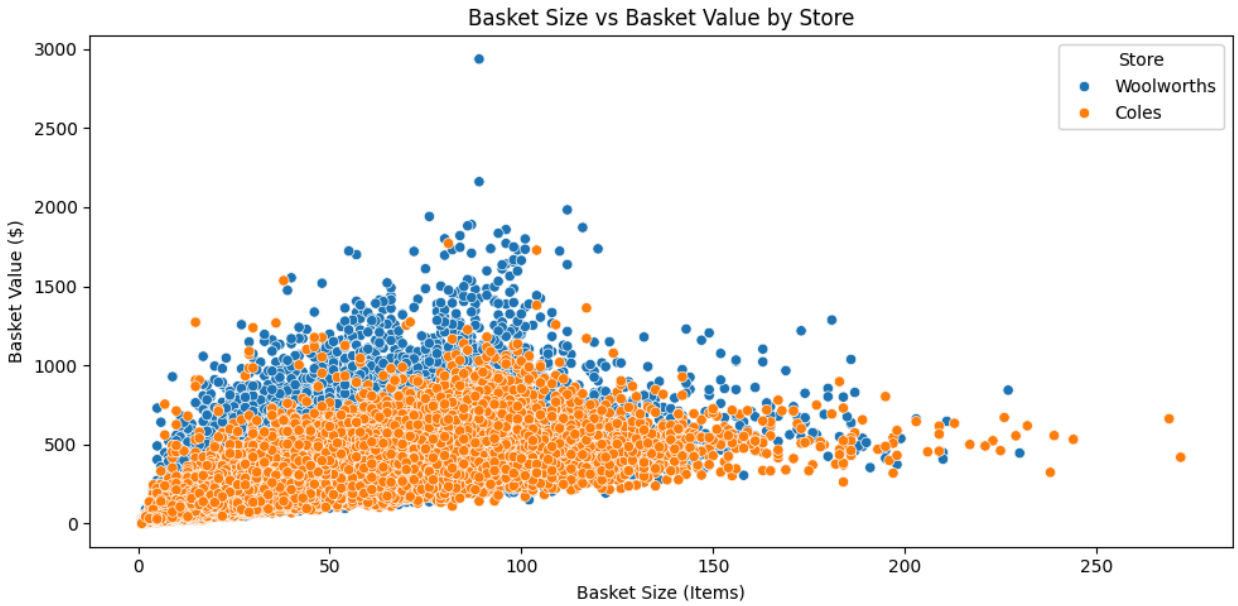


Graph Analysis:

- Each gender group (Female, Male, Other) shows **near-identical usage counts** for all payment types across cities.
- No gender-specific or city-specific preference emerges—**payment distribution is flat and balanced**.
- Indicates **random or fixed-proportion assignment** of payment methods, not behavior-driven.
- Highlights a core weakness of synthetic data: **lack of variability or realistic patterns**.
- Not suitable for tasks involving **customer segmentation**, demographic trends, or behavioral modeling.

This grouped bar chart breaks down **payment type usage by city and gender** (Female, Male, and Other). Across all combinations of gender and city, the distribution of payment methods, **Cash, Credit Card, EFTPOS, Gift Card, and Store Card** is nearly identical. This extreme uniformity is **unrealistic** and reinforces the limitations of synthetic data for behavioral insights. In reality, payment preferences would naturally vary across **genders, regions, and cities**, influenced by factors such as **income, age, and access to digital banking**. However, this dataset appears to have assigned payment types using a **uniform probability model**, ignoring any demographic or regional context. This kind of pattern-free consistency undermines the utility of the dataset for real-world **consumer behavior analysis**, and demonstrates that **synthetic data must be used cautiously**, particularly when analyzing human-centered variables.

Basket Size vs Basket Value by Store

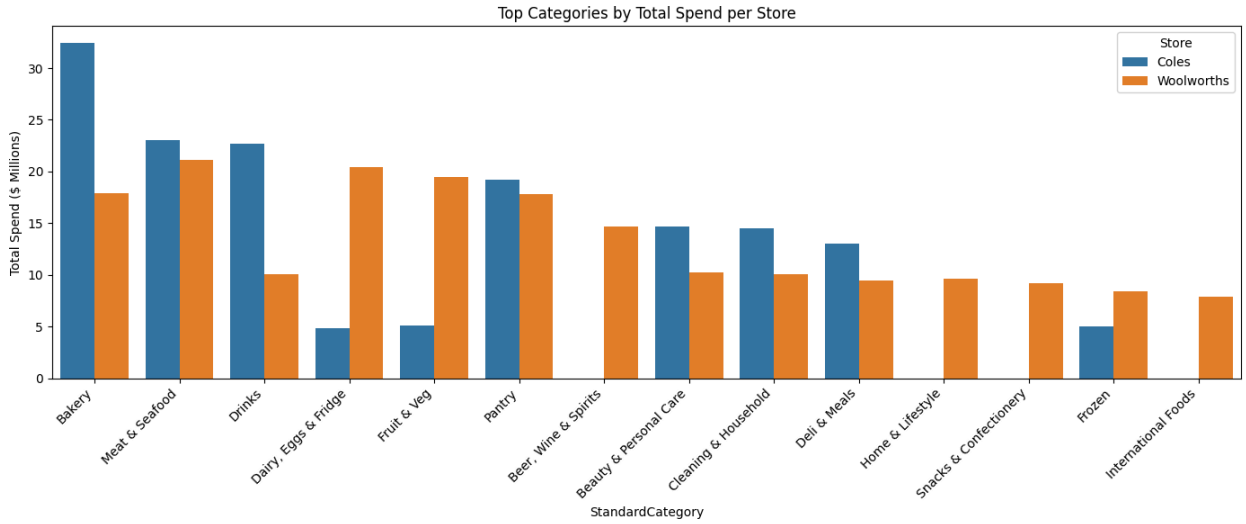


Graph Analysis

- Both Coles and Woolworths exhibit similar basket sizes, generally concentrated under 150 items.
- **Woolworths** transactions show significantly greater **variation in basket value**, with several purchases exceeding **$2000+**, whereas Coles rarely exceeds **$1000**.
- **Coles** data is tightly clustered in the **lower-left region**, indicating **smaller baskets and lower total value**.
- **Woolworths** displays a **wider spread** along both axes, suggesting both **larger baskets and higher-value purchases** occur more frequently.
- Outliers are more prominent in Woolworths, hinting at the presence of **bulk-buying or high-value item purchases**.

To deepen my analysis, I selected two new datasets—**customer_transactions_coles.csv** and **customer_transactions_woolies.csv**—sourced from the same synthetic dataset author on GitHub. I excluded IGA due to inconsistent formatting, which made comparisons impractical. I began by plotting **Basket Size vs Basket Value by Store**. This scatter plot revealed an interesting pattern: while both Coles and Woolworths had **comparable basket sizes**, the **basket values at Woolworths were far more variable**. Coles customers mostly made smaller, lower-value purchases, whereas Woolworths customers were more likely to have either **many items or higher-value transactions**. This discrepancy suggests differences in pricing strategy, product range, or synthetic modeling logic between the two datasets.

Top Categories by Total Spend per Store

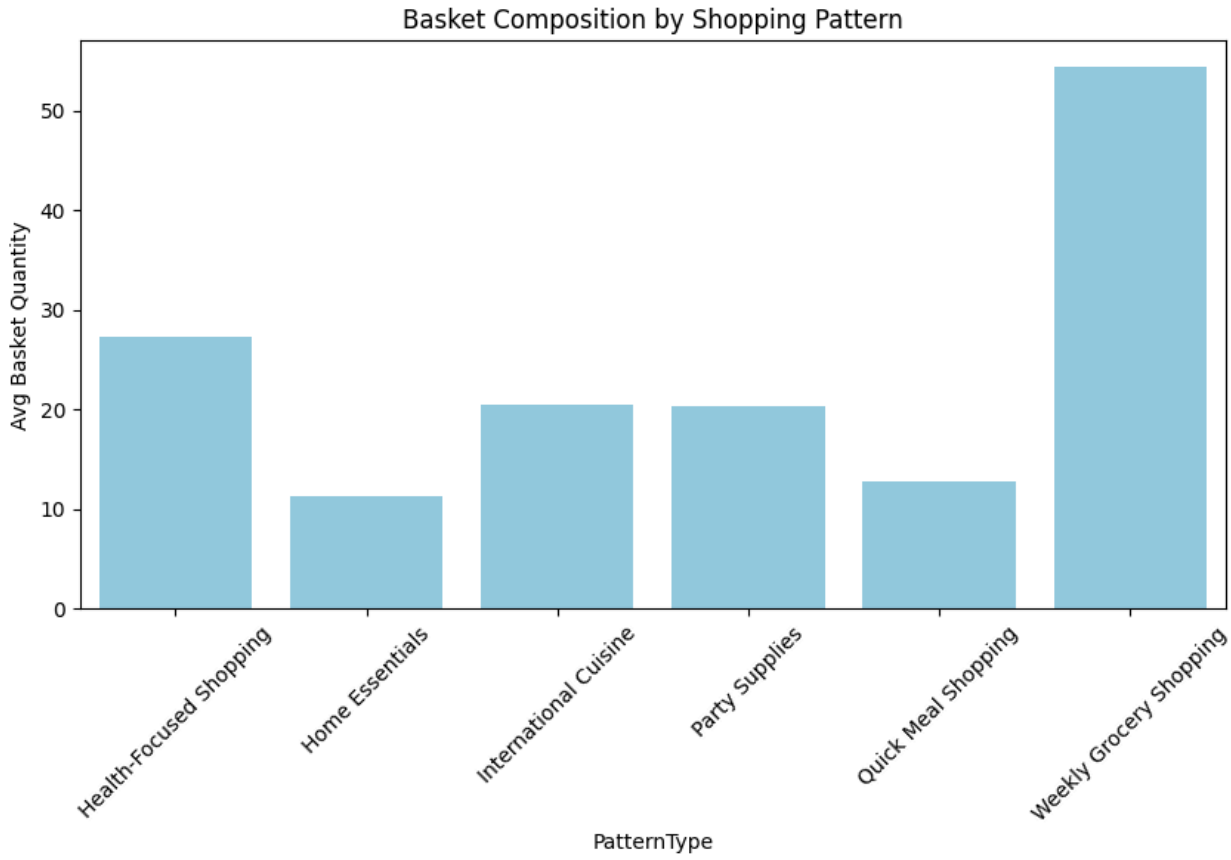


Graph Analysis

- **Coles** leads in categories like **Bakery, Drinks, Dairy & Eggs, and Meat & Seafood**, particularly Bakery, which surpasses **$30 million** in total spend.
- **Woolworths** dominates **Pantry, Fruit & Veg, and International Foods**, with consistently strong performance across a broader set of categories.
- Categories such as **Beauty & Personal Care**, **Cleaning & Household**, and **Deli & Meals** are more evenly matched, though slightly higher for Coles.
- Some categories like **Snacks & Confectionery** and **Home & Lifestyle** are only represented by Woolworths, suggesting **dataset coverage differences**.
- The overall trend shows **Coles excelling in staple and fresh food**, while **Woolworths shows strength in pantry goods and diverse specialty areas**.

To further compare Coles and Woolworths, I grouped and aligned their categories under a common structure to evaluate which departments contribute most to revenue. This bar chart displays **total spend per category by store**, revealing distinct spending patterns. **Coles** sees higher spending in everyday essentials such as **Bakery, Drinks, and Dairy**, with Bakery alone exceeding **$30 million** in sales. In contrast, **Woolworths** generates more revenue in **Pantry, Fruit & Veg, and International Foods**, suggesting a broader or more premium product range in those areas. The data implies that while Coles captures high-frequency, staple-driven transactions, Woolworths may benefit more from **variety and basket depth**, aligning with earlier findings from the basket value analysis.

Basket Composition by Shopping Pattern



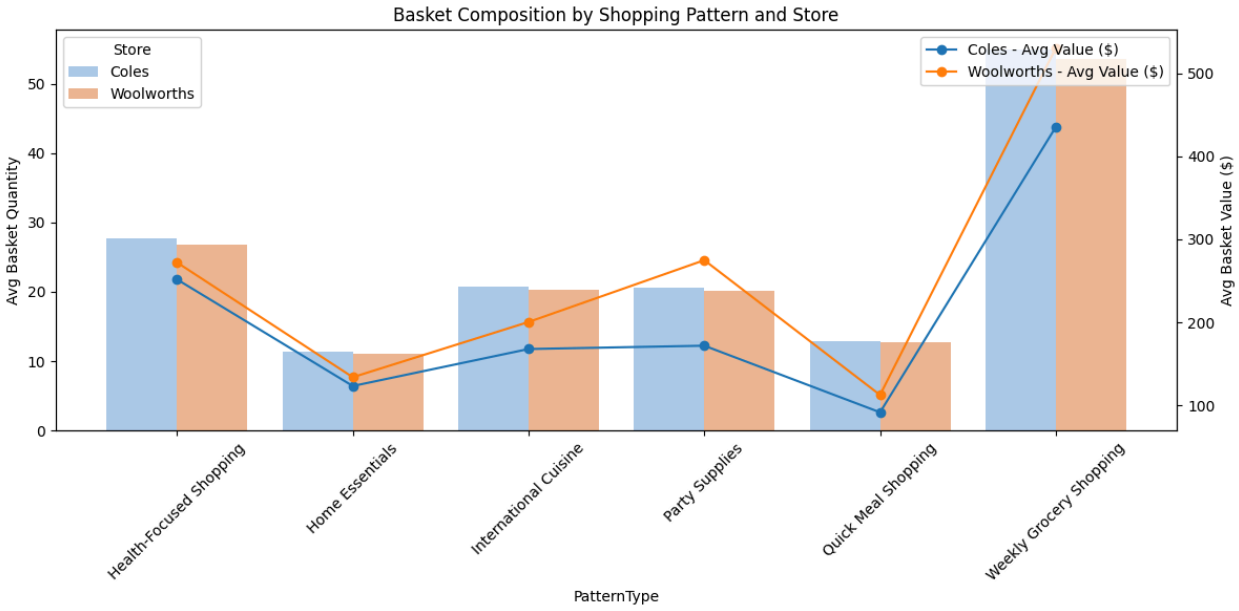
Graph Analysis

- **Weekly Grocery Shopping** has the highest average basket quantity by a large margin, exceeding **50 items per basket**.
- All other shopping patterns, including **Health-Focused**, **Party Supplies**, and **International Cuisine**, have basket sizes below **30 items**.
- **Home Essentials** and **Quick Meal Shopping** have the **smallest baskets**, averaging under **15 items**.
- The clear disparity in basket size suggests that **routine, high-volume grocery trips dominate**, while other patterns represent more focused, lower-volume purchases.

To better understand how different types of shopping behavior impact basket size, I analyzed **basket composition by shopping pattern**. The results show a striking dominance of **Weekly Grocery Shopping**, with an average basket size of over **50 items**—far surpassing every other

category. In comparison, categories like **Health-Focused Shopping**, **Party Supplies**, and **International Cuisine** average between **20–30 items**, while **Home Essentials** and **Quick Meal Shopping** are even smaller. This suggests that for both **Coles and Woolworths**, the core customer use case is **routine, high-volume grocery trips**, rather than specialty or one-off purchases. This trend reinforces the brands' positions as **general household grocery providers**, with infrastructure and stock tailored to serve **weekly family or household restocks**.

Basket Composition by Shopping Pattern and Store



Graph Analysis

- **Basket quantity** is nearly identical between Coles and Woolworths across all shopping patterns, with **Weekly Grocery Shopping** leading significantly.
- **Coles** shows consistently **higher average basket value** than Woolworths in most categories, particularly in **Health-Focused Shopping**, **Party Supplies**, and **Weekly Grocery Shopping**.
- Both stores have the **lowest basket value and quantity** in **Quick Meal Shopping**, suggesting smaller, lower-spend trips.
- The gap in value is most noticeable in **Party Supplies**, where Woolworths has a visibly higher quantity but Coles leads in value.
- **Overall pattern alignment** suggests similar customer behavior and product offerings, with slight pricing or purchase value differences.

To compare shopping behavior between **Coles and Woolworths**, I examined **basket composition by shopping pattern**, measuring both **average basket quantity** and **average basket value**. The results show that both stores exhibit very **similar purchasing patterns**, with nearly identical basket sizes across categories. However, **Coles consistently shows higher average basket values** in key categories such as **Health-Focused Shopping, Party Supplies,**

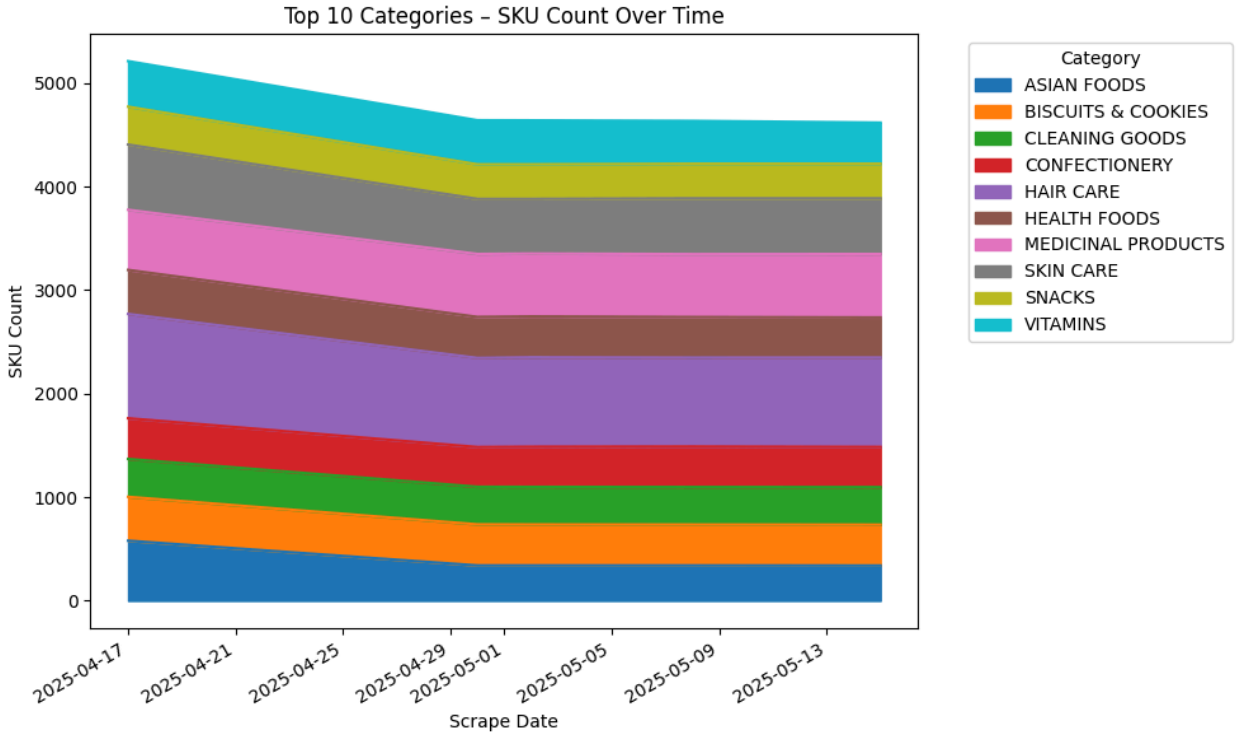
and Weekly Grocery Shopping. This suggests that Coles customers may either purchase more premium items or that Coles assigns **higher average prices** in the synthetic data. The consistency in quantity implies **aligned shopper intent**, while the variance in value highlights subtle differences in **spending behavior or pricing structure**.

Scraped Data - Coles

To address the issue of synthetic data, I used scraped data from DiscountMate's MongoDB. There I was able to collect scraped data from a variety of sources, however only files with the ending Coles_All were used as there were 6 files taken over the course of 1 month and several days. This was the most comprehensive series of files in the scraped data folder. While it was not adequate enough considering that the scrapes took place over a very short time span, 1 month was far longer than any alternative, many of which were several days at most.

The scraped data was the coles catalog, which included the prices of all items in the coles catalog. Due to this, changes in prices were noticed, either due to discounts or other measures which reduced or increased overall price.

SKU Count Over Time



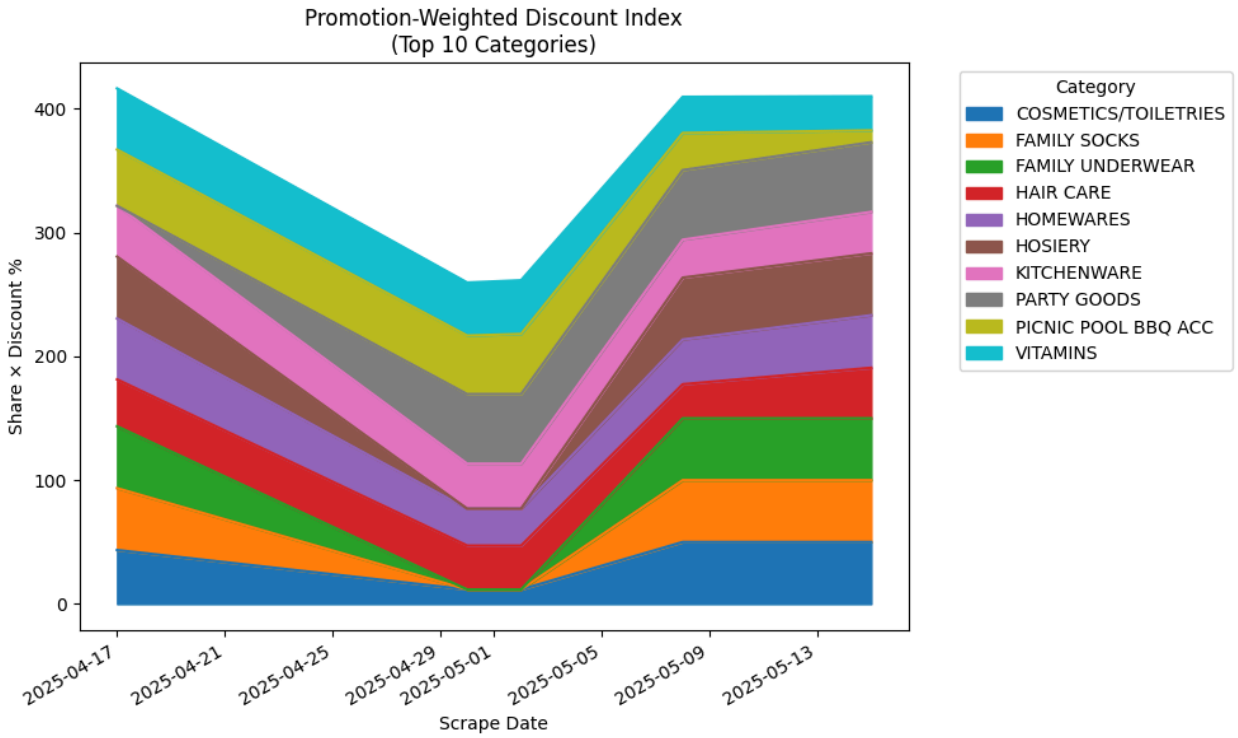
Graph Analysis

- The graph displays a **gradual decline in total SKU count** across all top 10 categories from April 17 to May 13, 2025.
- **Hair Care**, **Medicinal Products**, and **Skin Care** hold the largest share of SKUs throughout the time period.
- Categories like **Asian Foods**, **Biscuits & Cookies**, and **Vitamins** show the **steepest declines**, indicating potential stock reductions or de-prioritization.
- Between **April 25 and May 1**, the graph shows the **most significant drop**, after which SKU counts stabilize.
- The trend suggests **product range contraction**, either due to supply chain adjustments, assortment pruning, or temporary delistings.

This area chart tracks the **SKU count over time** for the top 10 product categories, offering a clear view of assortment trends between mid-April and mid-May 2025. A steady decrease in total SKU availability is evident across nearly all categories, with the most dramatic drop

occurring around the end of April. While categories like **Hair Care**, **Medicinal Products**, and **Skin Care** continue to dominate in count, others—particularly **Asian Foods**, **Biscuits & Cookies**, and **Vitamins**—show notable reductions. This pattern could reflect changes in supplier availability, seasonal shelf resets, or strategic reductions in product variety. After the initial decline, the SKU count plateaus, suggesting the system may have stabilized with a leaner product offering.

Promotion-Weighted Discount Index



Graph Analysis

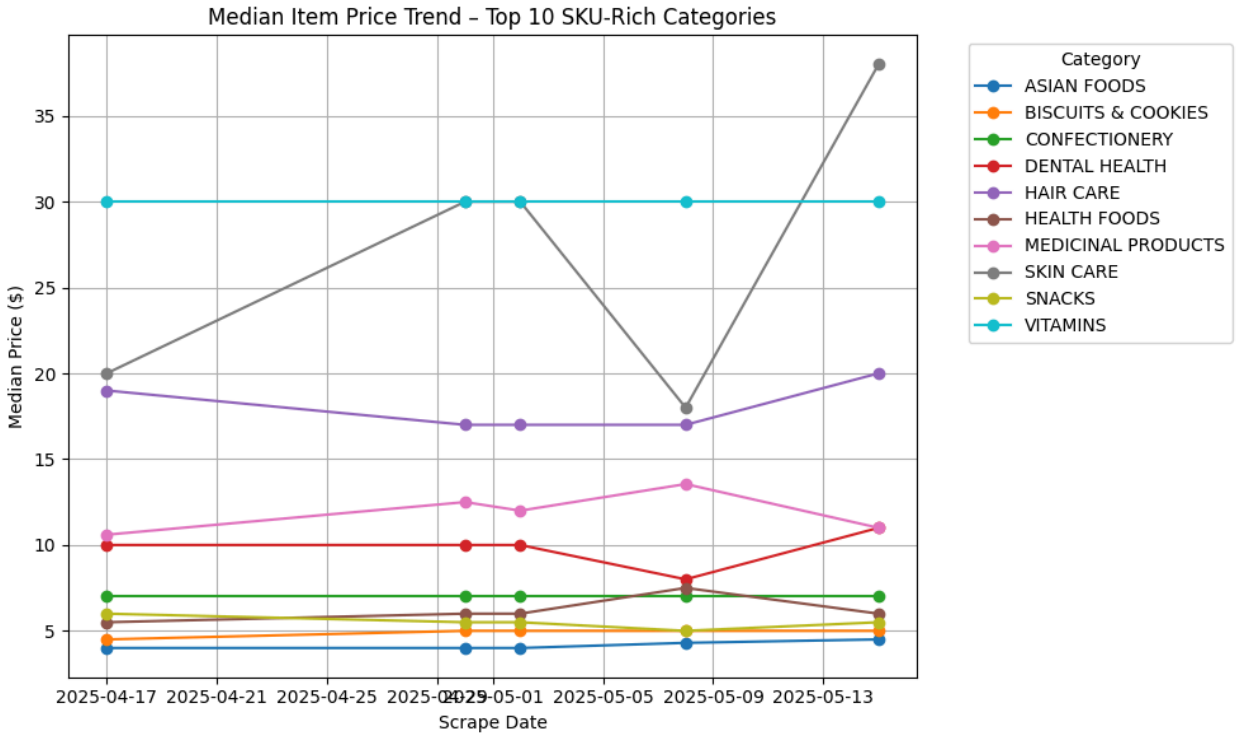
- The **Promotion-Weighted Discount Index** (Share × Discount %) dropped sharply across all top 10 categories between **April 17 and May 1**, hitting a low in early May.
- Categories like **Cosmetics/Toiletries**, **Family Socks**, and **Hair Care** show strong promotional presence throughout but dip significantly before rebounding.
- **Party Goods** and **Picnic Pool BBQ Accessories** contribute more prominently in the recovery phase, especially after May 5.
- The **rebound in promotional activity** after May 5 suggests the end of a low-discount promotional period, possibly following a seasonal or campaign lull.
- Despite the dip, **Vitamins**, **Hair Care**, and **Kitchenware** maintain a relatively steady discount presence over the entire time period.

This stacked area chart visualizes the **promotion-weighted discount index** for the top 10 product categories, reflecting how much discount activity each category experiences over time.

The index is calculated by combining each category's promotional share with the average discount offered, giving a sense of both intensity and scope. A significant drop in total promotional volume is seen between **April 17 and May 1**, reaching a low in early May. This suggests a **brief pause in major promotional activity**, perhaps between campaign cycles. However, discounts recover quickly by mid-May, with a diverse mix of categories—particularly **Party Goods**, **Picnic & Pool Accessories**, and **Cosmetics/Toiletries**—regaining visibility. The temporary dip and rebound hint at **seasonal campaign pacing**, where discount intensity fluctuates strategically. Despite the changes, **Vitamins and Hair Care** remained steadily promoted, showing their consistent role in promotional strategies.

#	Category	otion-Weighted Discount I
#1	PARTY GOODS	56.25
#2	FAMILY UNDERWEAR	50
#3	HOSIERY	50
#4	FAMILY SOCKS	50
#5	VITAMINS	38.487008
#6	HOMEWARES	37.500308
#7	PICNIC POOL BBQ ACC	36.171345
#8	KITCHENWARE	35.559487
#9	HAIR CARE	35.434212
#10	COSMETICS/TOILETRIES	33.37727
#11	BOXED CHOCOLATES	32.909504
#12	EASTER	30.797918
#13	DISHWASHING	27.99989
#14	LAUNDRY	26.437589
#15	SEASONAL EVENTS	24.22773
#16	ELECTRICAL	23.605058
#17	SKIN CARE	23.556302
#18	SOAPS & BODY WASH	22.869727
#19	HOUSEHOLD APPLIANCES	22.720322
#20	CONVENIENCE MEALS	22.199597

Median Item Price Trend



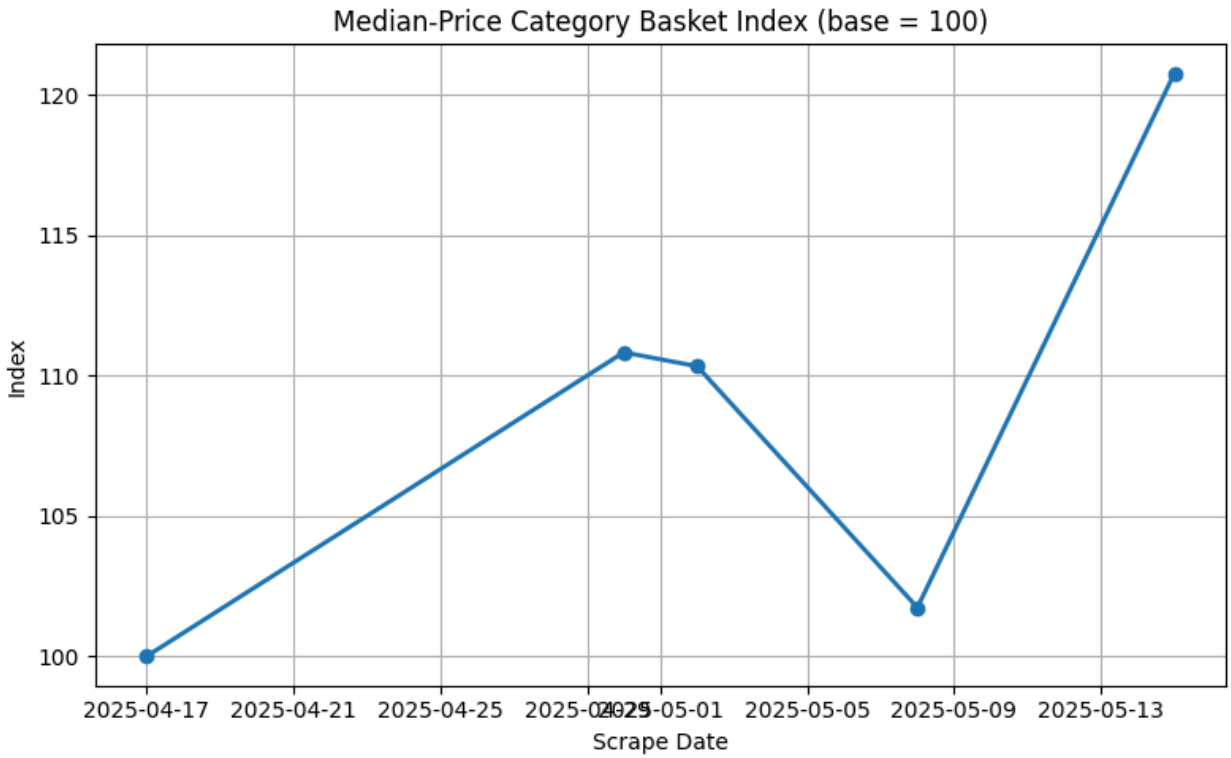
Graph Analysis

- **Skin Care** shows the most volatility, with median price jumping from **$20 to $30**, briefly dipping to **$18**, then peaking at **$38** by May 13.
- **Vitamins** maintain a **constant median price of $30**, showing no fluctuation across all dates.
- **Dental Health** remains stable at **$10** until a temporary drop to **$8**, then recovers to **$11**.
- **Hair Care** slightly decreases from **$19 to $17**, then returns to **$20**.
- **Asian Foods, Biscuits & Cookies, Confectionery, and Snacks** show minimal or no price movement, remaining under **$7**.
- **Medicinal Products** gradually increase in price before stabilizing, suggesting slow but consistent pricing adjustments.

This line chart illustrates **median item price trends** for the top 10 SKU-rich categories between April 17 and May 13, 2025. The most dramatic pricing behavior is observed in **Skin Care**, which rises sharply by **90%**, ending at **$38**, a sign of either sudden product mix shifts or pricing strategy changes. In contrast, **Vitamins** maintain a **fixed price of $30** throughout the period, indicating stable product pricing. Categories like **Asian Foods**, **Biscuits & Cookies**, and **Snacks** remain stable with very low median prices, reinforcing their role as low-cost, high-frequency goods. **Hair Care** and **Medicinal Products** show minor but meaningful movement, with **Medicinal Products** steadily increasing and **Hair Care** slightly rebounding after a dip. Overall, the plot reveals a mix of **stable essentials and price-sensitive categories**, offering insight into which segments are subject to more frequent price changes or promotional influence.

Category	First Price ($)	Last Price ($)	% Change
SKIN CARE	20	38	90
ASIAN FOODS	4	4.5	12.5
BISCUITS & COOKIES	4.5	5	11.111111
DENTAL HEALTH	10	11	10
HEALTH FOODS	5.5	6	9.090909
HAIR CARE	19	20	5.263158
MEDICINAL PRODUCTS	10.6	11	3.773585
CONFECTIONERY	7	7	0
VITAMINS	30	30	0
SNACKS	6	5.5	-8.333333

Median Price Category Basket Index

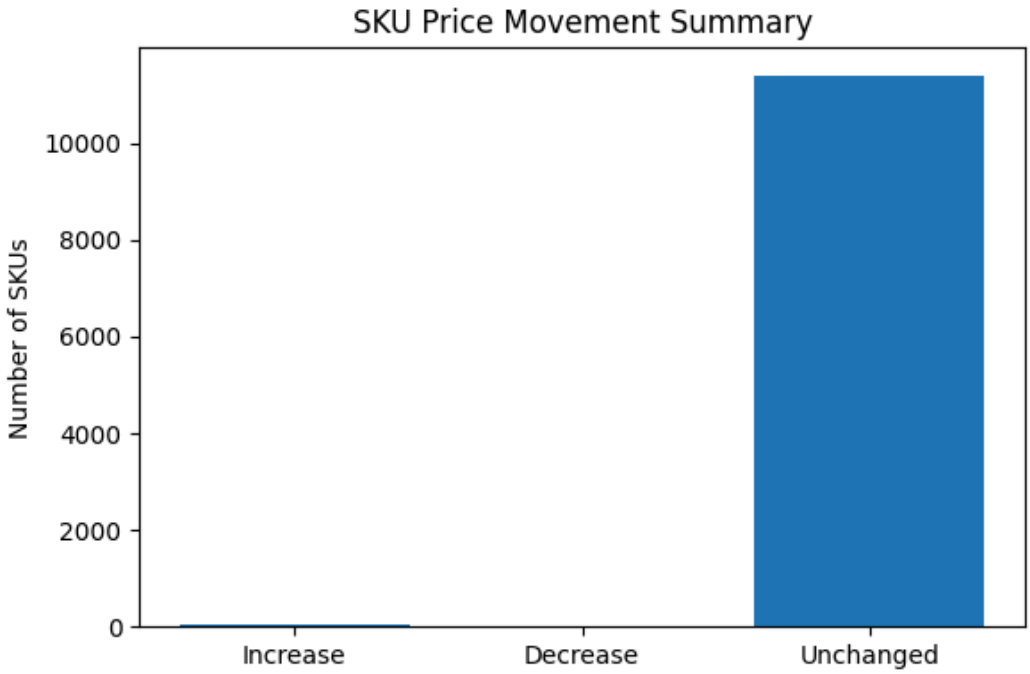


Graph Analysis

- The **Median-Price Category Basket Index**, which uses April 17 as the base (100), shows a clear upward trend overall.
- The index rises steadily to **111** by April 30, indicating a **11% increase** in median-category prices.
- A **short-term decline** follows, dropping to just above **101** on May 5, suggesting a temporary reduction or promotion.
- By May 13, the index peaks at **121**, marking a **21% total increase** from the baseline.
- The curve reflects a **volatile pricing environment** with a strong surge in the final scraping period.

This line chart tracks the **price index for a median-priced category basket**, normalized to a base value of 100 on April 17. Initially, the index climbs steadily, reaching 111 by the end of April, suggesting a gradual increase in overall pricing. A brief dip to 101 on May 5 may represent promotional activity, clearance, or temporary price normalization. However, this is followed by a sharp rebound, with the index peaking at **121** by May 13—indicating a **21% rise in median category prices** over less than a month. This suggests either a substantial **shift in product mix**, **seasonal price adjustments**, or possible **supply-side pressures** affecting median-priced items. The price volatility reflected in this chart aligns with earlier observations from category-specific trends and underlines the importance of ongoing price monitoring in key segments.

SKU Price Movement Overview

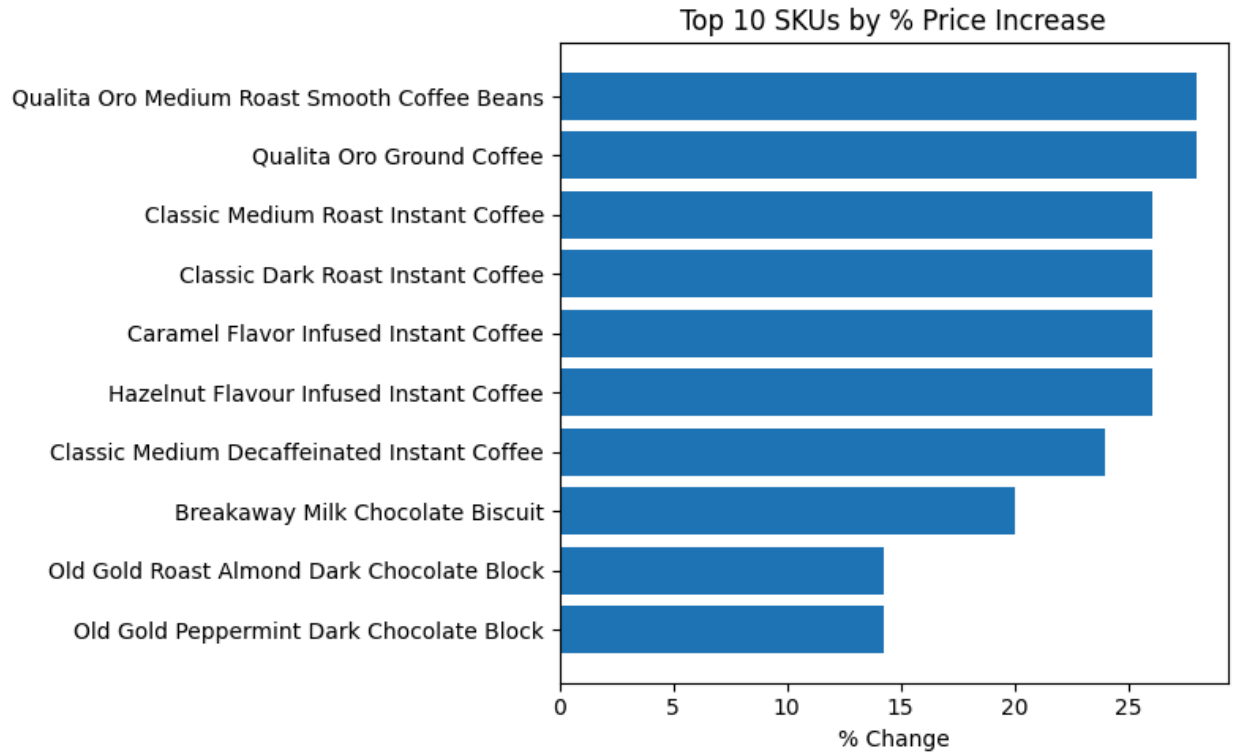


Graph Analysis

- The vast majority of SKUs—**over 11,000**—show **no price change** over the observed period.
- A **very small number** of SKUs experienced a **price increase**, while **none** registered a decrease.
- The chart shows an **extreme skew** toward static pricing, with minimal dynamic price behavior.
- This suggests a **high degree of price rigidity**, at least during the time frame of analysis.

This bar chart summarizes SKU-level price movements across the dataset. It reveals that nearly all items—**over 11,000 SKUs**—had **unchanged prices** across the entire observation window. Only a **handful of SKUs** experienced price increases, and **none recorded a decrease**. This pattern of near-total price stasis may point to a few different explanations: the dataset may represent a narrow time slice, the synthetic data may have been generated without realistic pricing fluctuations, or the product range remained largely unaffected by promotions or cost changes during the observed period. Regardless of the cause, this visual underscores a lack of meaningful SKU-level price dynamics, limiting the utility of this dataset for inflation tracking or elasticity modeling.

Top 10 SKUs by Percentage Price Increase



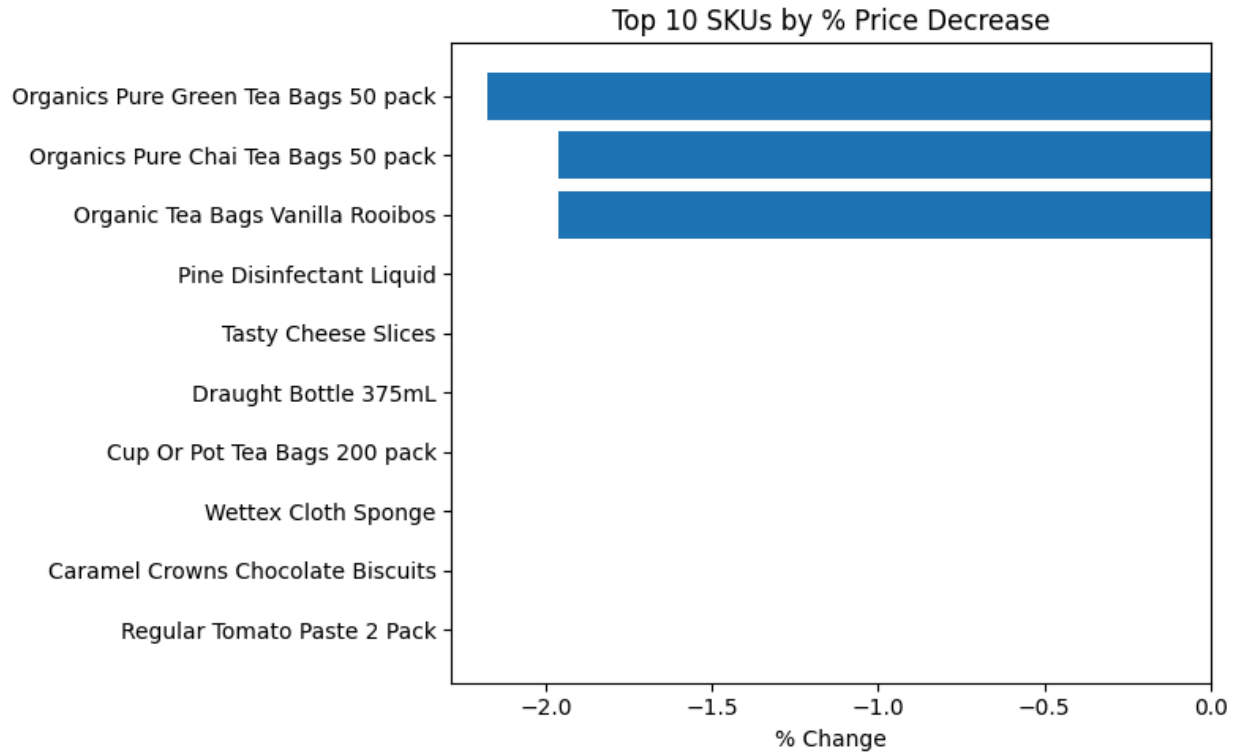
Graph Analysis

- The SKUs with the highest price increases are **heavily dominated by coffee products**, with the top 7 all being variations of **instant or ground coffee**.
- **Qualita Oro Medium Roast Smooth Coffee Beans** and **Qualita Oro Ground Coffee** top the chart with nearly **27% increases**.
- The remaining SKUs are primarily **chocolate and biscuit products**, with **Old Gold** and **Breakaway** brands appearing.
- All listed products experienced increases of **at least 14%**, signaling targeted inflation or pricing adjustments in specific high-demand categories.

This horizontal bar chart displays the **top 10 SKUs by percentage price increase**, offering a more granular view of pricing shifts at the product level. The majority of the highest increases occurred in the **coffee segment**, with multiple entries under brands like **Qualita Oro** and **Classic Roast**, all exceeding **20% increases**. This trend may reflect **rising wholesale coffee prices**, **supply constraints**, or strategic pricing decisions targeting daily staples. The presence

of **chocolate and biscuit SKUs**, particularly **Old Gold** and **Breakaway**, suggests similar pricing behavior in indulgent or branded snack segments. These increases contrast sharply with the general SKU movement summary, where almost all prices remained unchanged, highlighting that **only a very specific subset of SKUs is driving any upward price momentum**.

Top 10 SKUs by Percentage Price Decrease



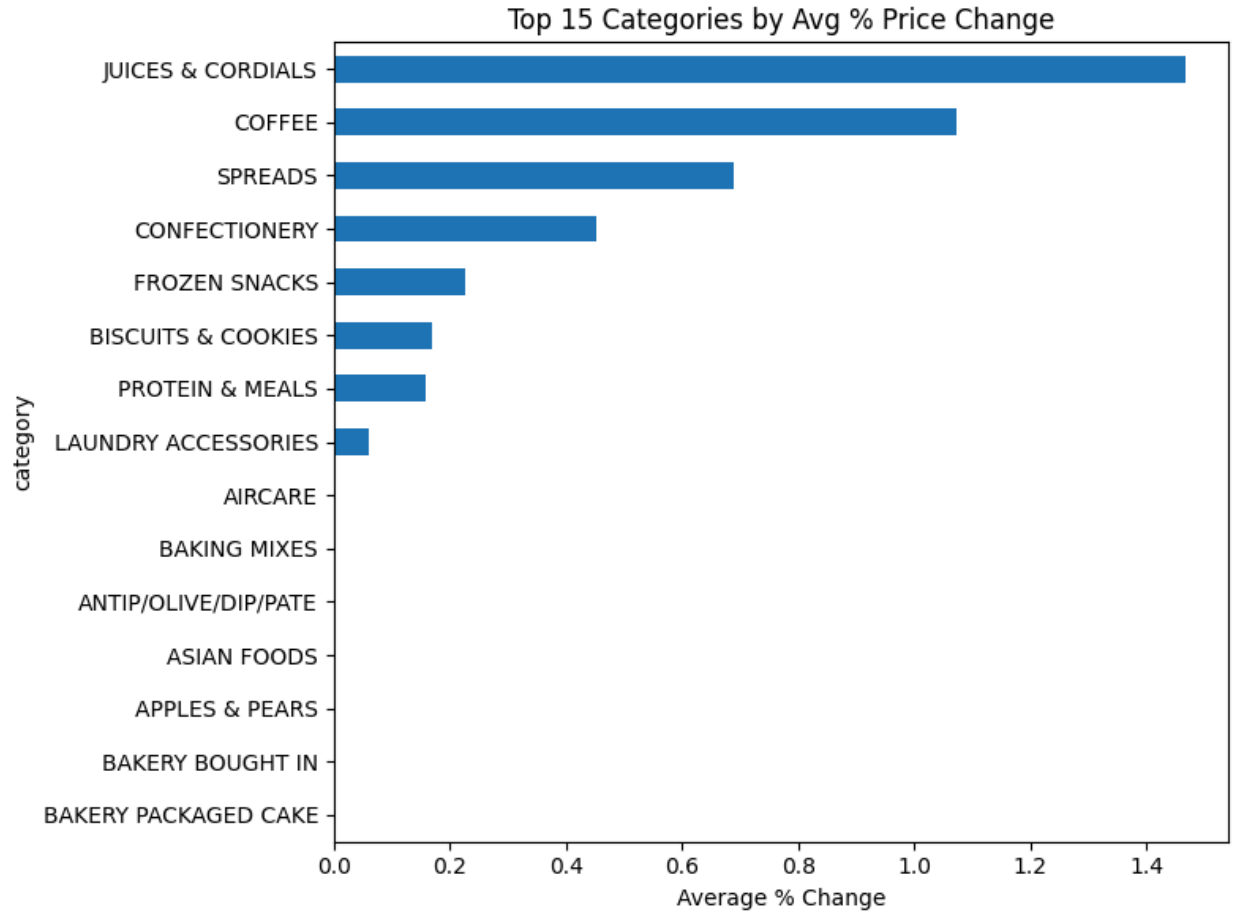
Graph Analysis

- The price decreases are **very small in magnitude**, with the largest being just above **2%**.
- The top three SKUs are all **organic tea products**, indicating targeted price cuts in that niche.
- The rest of the list includes a **diverse mix** of household and grocery items: **disinfectant, cheese slices, tomato paste, and chocolate biscuits**.
- Most of the SKUs on this list reflect **modest discounting**, possibly temporary or promotional in nature.
- Compared to the price increase list, the **extent and depth of price decreases are minimal**, confirming an **overall upward pressure on prices**.

This chart highlights the **top 10 SKUs by percentage price decrease**, providing insight into where rare downward price movements occurred. The most notable trend is that the **top three**

SKUs are all **organic tea products**, such as **green tea, chai, and vanilla rooibos**, each dropping by just over **2%**. Beyond that, the decreases become even more marginal. Products like **pine disinfectant, cheese slices, tomato paste**, and **chocolate biscuits** make up the rest of the list but show declines that are likely **too small to reflect significant promotional strategy**. This supports a broader conclusion that **price reductions in the dataset are limited in both frequency and intensity**, and that the dataset skews heavily toward **price stability or increase**, with very few items showing measurable decline.

Top 15 Categories by Average Percentage Price Change



Graph Analysis

- **Juices & Cordials** lead all categories with the highest average percentage price increase, exceeding **1.4%**, suggesting strong inflation or reformulation-driven pricing.
- **Coffee** follows closely behind, reinforcing earlier SKU-level findings of consistent price hikes in caffeinated products.
- **Spreads** and **Confectionery** also show significant increases, both above **0.4%**, indicating rising costs in sweet and breakfast-related items.
- The rest of the list consists of moderate risers like **Frozen Snacks**, **Biscuits & Cookies**, and **Protein & Meals**, with increases under **0.3%**.

- The lower portion of the chart includes categories like **Aircare**, **Asian Foods**, and **Bakery Packaged Cake**, which show **negligible to zero average price movement**.

This horizontal bar chart ranks the **top 15 categories by average percentage price change**, helping identify which segments experienced the most pricing pressure. The standout is **Juices & Cordials**, which saw the steepest average increase, pointing to possible supply chain shifts or new product introductions. The **Coffee** category maintains its upward trend, matching earlier observations from SKU-level analysis. Other food staples like **Spreads, Confectionery, and Frozen Snacks** also recorded price hikes, potentially reflecting broader **cost-push inflation** across processed goods. Notably, many categories toward the bottom show **zero price movement**, reaffirming the trend that **price increases are highly concentrated in select segments**, while most others remain static or inert—particularly in synthetic datasets where real-world variability may be suppressed.

Conclusions

The exploratory analysis of the synthetic retail datasets reveals several broad themes and important caveats:

1. **Pantry as a Focal Category**

    - *Why Pantry?* It records the highest unit sales, making it a natural proxy for core grocery demand.

    - *Temporal Signals* A modest but clear **weekly cycle** (ACF ≈ 0.16 at lag 7) confirms that most Pantry purchasing follows a weekly rhythm. Beyond this, trend decomposition shows only mild, short-lived rises (e.g., November and late February) and **weak seasonality**; noise and one-off promotions dominate day-to-day fluctuations.

2. **Synthetic Data Limitations**

    - **Demographic and payment variables** (gender splits, payment type usage) are distributed almost perfectly evenly across locations—an implausible pattern in genuine retail data.

    - **Total-sales heatmaps** and **location–store comparisons** show less than 10 % spread between “best” and “worst” performers—again, unrealistically tight.

    - Such uniformity is useful for demonstration or load-testing dashboards, but it impedes any realistic behavioural or segmentation work.

3. **Coles vs Woolworths—Key Contrasts**

    - **Basket profiles**: Basket sizes are similar, yet Woolworths exhibits far greater **basket-value dispersion**, including the dataset’s largest outliers (> $2 000).

    - **Category spend**: Coles leads in staple, high-frequency lines (Bakery, Drinks, Dairy & Eggs), whereas Woolworths excels in Pantry, Fruit & Veg, and International Foods—implying a broader or more premium range.

    - **Basket composition by shopping pattern**: Quantities align closely across stores, but **Coles’ average basket value is consistently higher**, hinting at higher per-item prices or premium product mixes in the synthetic specification.

4. **Shopping-Pattern Insights**

- **Weekly Grocery Shopping** dwarfs every other pattern, averaging > 50 items per basket; all other segments (Health-Focused, Party Supplies, etc.) remain well below half that size.

- This mirrors real-world positioning: both chains function primarily as weekly household-stock-up destinations rather than specialty retailers.

5. **Overall Takeaways**

    - The datasets successfully illustrate *methodology*—time-series decomposition, basket analysis, category ranking, and cross-store comparisons—but their synthetic nature drives **artificial uniformity** in payment behaviour, demographics, and regional performance.

    - Any operational or marketing decisions drawn from these data should therefore be **treated with caution** or validated against real transaction records.

    - For future synthetic-data generation, injecting realistic demographic skews, price dispersion, and regional variability would yield richer, more actionable analysis.

In short, the analysis confirms a weekly purchasing cadence in core grocery lines and highlights meaningful (albeit synthetic) differences between Coles and Woolworths basket economics. Yet the pronounced uniformity across demographic and payment dimensions underscores the **critical need for real-world data** when the goal is genuine customer or locality insight.

Disclaimer: This report has been beautified by ChatGPT